JOURNAL OF PURE AND APPLIED MICROBIOLOGY, Nov 2015. Vol. **9**(Spl. Edn. 2), p. 345-349

An Enhanced Method for Disease Prediction using Ordinal Classification-APUOC

L. Mary Gladence[1], T. Ravi[2] and Y. Mistica Dhas[3]

[1]Research Scholar, Sathyabama University, Chennai. India.
[2]Madanapalle Institute of Technology& science, P.B.NO:14, Angallu, Madanapallee - 517 325, India.
[3]Department of Information Technology, Sathyabama University, Chennai, India.

(Received: 26 July 2015; accepted: 01 October 2015)

Medical Management collects vast amount of data in which retrieving the data in the form of useful as well as not useful data becomes important task. Discovery of this hidden information, Patterns, Relationship often goes undeveloped.Most influential concept in the Data Mining is Classification which has become the most significant method while predicting Disease. Existing work is done based on Nominal Classification which has produced the results in the form of Yes/No.This research uses Ordinal classification concept to produce accurate results. A Novel Technique which proposed in this work can remedy the drawbacks which are faced during earlier research. Proposed work developed a new technique called APUOC. Here, data set is classified based on random threshold value'¸'. Based on these the distance between the different class labels are predicted by All PairsDistance Calculation using Ordinal Classification technique. By doing in this manner there won't be mismatch while predicting the disease. Taking all these into account, new trained data is in the form of outcome of the proposed work. Here, with the knowledge of proposed work testing data is tested with new trained data sets and the results are predicted. This proposed work is analyzed with existing algorithm such as Kernel Discriminant Analysis, Logistic regression, Classification via Regression, Multiclass classifier etc.,

Key words: Data collection; Ordinal Classification; Distance Calculation; Projection; Disease Prediction;Data Mining.

In, Nominal Classification features which are similar are grouped in the same class and remaing are grouped under different class and it will give the output as "1" or "0". Noticing in this aspect there is a chance of misclassification, especially in disease prediction.To overcome this drawback which are presented in the Nominal classification Proposed work is evaluated using Ordinal Classification[1] to[13]. In Ordinal Classification, Ordinal class labels are assigned for each data and the datas are grouped in a more appropriate way, i.e it is categorized in to different classes.This is useful if we are considering disease prediction, instead of getting the result as "Yes" or "No" class label results are in the form of "Severe", "Moderate", "slight". This shows that, Ordinal regression avoids inconsistent data while comparing it with Nominal Classification.

Ordinal classification predicts predictor variable and response variable instead of predicting the categorical class label. From this we observed thatRegression Problem is mapped to a real data while classification problems are mapped to an unordered set of classes.Here, main goal of the work is done in training phase itself. The essential idea behind this paper is to perform the projection model based on class distribution obtained from the distance measures using the concept of All

* To whom all correspondence should be addressed.
E-mail:lgladence@gmail.com

Pairs Distance Calculation. The proposed work is analyzedwith following standard methods such as kernel discriminant analysis, support vector machines, Logistic regression,Classification viaRegression, Multiclass Classifier,J48.

WojciechKotBowski et.al[2] proposed an ordinal classification along with monotonicity limitation. It varies from classification by the following factors such as Knowing background knowledge about ordering classes, ordered attributes, monotonic constraints between an object on the attributes and its class description. We can even define this by using another way called class label, i.e.Whenever input variables increases output variable should not decrease.Here an analysis of nonparametric approach is done using statistical point of view. M.Perez-Ortiz et al[8], developed a method for reducing the dependency between classifiers. This method can be used as an added core to maximize the diversity, develop the essential combination of rules. This method can be used with any threshold value as a base classifier. Although, there are many classifiers due to many classes each single model differs from the remaining ones as ordinal ranks are used,i.e. differentiating each and every class with one another. This work is based on disintegrating ordinal regression problem into simpler classification tasks in which order of each and every detail is expressed.

ShereenFouadet et.al[4] proposed Learning Using privileged Information (LUPI). Supervised learning is improved in the presence of privileged information. Here, information is available only in the training phase not in testing phase. This Novel Learning Methodology is expressed to incorporate privileged information in ordinal classification tasks,whenever natural order is followed. By changing the global metric in the input space, based on distance relations revealed by the privileged information is used to justify this work. Experiments shows that by integrating privileged information via the proposed ordinal-based metric learning can upgrade the ordinal classification performance.

Yi Liu and Yuan F. Zheng et.al[11] uses Support Vector Machine concept, but it has the major disadvantage of ignoring the overall information about the data and its high computational complexity.To avoid these problem Proposedmethod, consider entire data points using All Pairs Distance Calculation using Ordinal Classification Method. After finding the results using Proposed work, these are compared with Kernal Discriminant Learning, Support Vector Regression, Logistic Regression. The Objective of this work is to reduce misclassification error by explicitly understand the data in the dataset during the training itself. So, testing data space (i.e) input space is mapped to already trained dataset value of considering the behaviors of that particular dataset. For example, if we aregoing to classify the heart disease related data,then the behavior of that particular dataset is explicitly known instead of direct method using one against all method. So if test data is applied, our algorithm will understand it explicitly and reduce the misclassification error.

All Pairs Distance Calculation Using Ordinal Classification[APUOC]

Initially the untrained dataset is trained based on All Pairs Distance Calculation using Ordinal Classification. “Fig.1” represents the architecture of the entire work. Here disease data set which are extracted from uci repository is taken as input. Each class is separated in the form of hyperline based on threshold values(threshold value is user defined one). Based on threshold value class labelas are framed. After separating data into different class label, minimum distance between each and every data should be found to validate whether identified class labels are correct or not.After finding minimum distance between each data points,dissimilar data points are adjusted with other classes.Now the result is in the form of new trained data set. This new trained data set is tested with testing data set to find out the disease

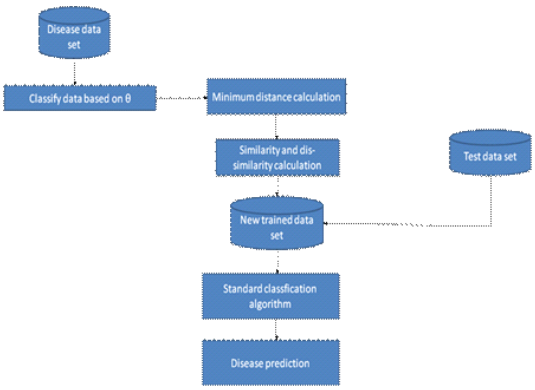

Fig. 1. Architecture of APUOC Method

prediction. Discovered results are tested with exixting algorithms such as KDL,LR,CVR etc.,

Pseudo code for the proposed [APUOC]

Training Phase(Phase-I)

Input: Training dataset

Output: Trained dataset

1.Load Data set.

2.Separate the classes based on threshold value' ¸'.

3.Find minimum distance for each data point within the class.

4.Compare each class data points with other class data points.

5.Adjust the data point based on minimum distance Calculation.

6.Repeat the process till all data points converges in their Classes.

Prediction(Phase II)

Input: Regression (Rg), threshold ($\boldsymbol{\theta}$), new trained

dataset, test data set

Output: Predicted value

1. Predict the latent variable value using the Regression(Rg) .

2. Map the regression (Rg),to the corresponding Class and find predicted value

4.Experiment Results

To validate the Proposed methodology some data set like HeartDisease, Contact-lenses, Thyroid,etc., are used. These datasets were extracted from uci repository. Table 1 shows the description of these, where number of patterns(N), number of attributes(K), number of classes(C) are taken. Initially, thesedatasets do not represent ordinal classification, but it represents regression. To evolve these regression into ordinal classification we have considered the desired result is categorized into five classes with equal frequency and these are referred from[2,3,4].

The Entire work is validated based on All Pairs Distance Calculation using Ordinal Classification and these results are compared using, Kernal Discriminant Analysis, Logistic Regression,Multi Class Classifier, Classification Via Regression etc..

Performance Measures

In this work Accuracy, Mean Absolute Error, Average Mean Absolute Error, Kendall's are used to validate the Proposed work. Accuracy is defined as how much percentage a given criteria guess the predicted attribute for the new data and it is determined using the equation(7).

$$acc = \frac{1}{m}\sum_{i=1}^{m} |y^{\propto}{}_{i} - y_i| \quad ...(7)$$

The Mean Absolute error is close to the Mean Squared Error. It uses absolute value inplace of squaring. Average of this is consider as Mean Absolute Value and it is measured using equation(8)

$$MAE = \frac{1}{m}\sum_{i=1}^{m} e(x_i) \quad ...(8)$$

Average Mean Absolute Error is measured using equation(9)

$$AMAE = \frac{1}{m}\sum_{m=1}^{m}(MAE_m) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{N_m} \quad ...(9)$$

Table 1. Characteristics of Datasets based on their classes

Dataset	N	K	C
Heart	270	13	5
Contact-lenses	18	6	3
Thyroid	161	6	3

Table 2. Measures considered for the Proposed APUOC and its Methods Compared

Dataset/ Measures	Accuracy	MAE	AMAE	Kendall's
Heart	0.851	0.375	0.352	0.887
Contact-lenses	0.721	0.299	0.512	0.650
Thyroid	0.963	0.017	0.061	0.875

Table 3. Comparision of APUOC's Accuracy with other Methods Compared

Dataset/Method	LR	APUOC	J48	CVR	MCC	KDA
Heart	0.562	0.851	0.792	*0.812*	0.739	0.742
Contact-lenses	0.694	*0.721*	0.733	0.719	0666	0.578
Thyroid	*0.963*	0.981	0.944	0.962	0.912	0.957

Table 4. Comparision of APUOC's Mean absolute Error with other Methods Compared

Dataset/Method	LR	APUOC	J48	CVR	MCC	KDA
Heart	0.256	0.375	0.352	0.887	0.245	*0.382*
Contact-lense	0.299	0.538	0.222	0.045	0.195	*0.352*
Thyroid	0.033	0.291	0.025	0.045	*0.067*	0.017

Table 5. Comparision of APUOC's Kendall's Measures with other Methods Compared

Dataset/Method	LR	APUOC	J48	CVR	MCC	KDA
Heart	0.752	0.423	0.725	*0.629*	0.726	0.887
Contact-lense	0.473	0.650	*0.428*	0.877	*0.715*	0.429
Thyroid	0.919	0.875	0.916	*0.877*	0.916	0.959

Mean Zero-one Error finds the misclassification rate and the rate of incorrect classified pattern is measured using equation (10)

MZE=1-acc ...(10)

After Evaluating measures such as Accuracy, Mean Absolute Error, Average Mean Absolute Error, Kendall's, results are displayed in Table 2.

Table 2 shows various performance measures such as Accuracy, Mean Absolute Error, Average Mean absolute Error, Kendall's measure of Proposed APUOC with different datasets. Table 3 shows about comparison between different methods like LR,J48,CVR,MCC,KDA with APUOC for different datasets based on Accuracy Values.Here Bold faced data represents maximum accuracy achieved for each data set and italic font represents second second highest accuracy achieved for each data set.Forexample,in Table 3 Heart Disease Dataset and Thyroid Data set has achieved maximum accuracy in proposed APUOC and second highest accuracy in Classification via Regression Method.

Similarly Table 4 and Table 5 are derived for different datasets based on Mean absolute Error and Kendall's Performance Metrics. In Table 4 APUOC has given least Mean Absolute Error for Heart Disease Data Set as well as Thyroid Data Set and second best result has been given by MultiClass Classifier and Classification via Regression Method.Similarly in Table 5 Kendall's Measures has given the best result using APUOC methof for Conrtact-lense as well as Thyroid data set. CVR has given the best result for Heart Disease Datset.

From these we can easily identify Proposed APUOC gave better result while compared with other methods. In addition to this, how values varies between various measures are also compared. However,Ordinal dataset gave better result in APUOC but in some other cases it gave second best values. For efficient comparision, we compared ordinal method result such as APUOC,LR,CVR as well as Nominal methods like J48,MCC

CONCLUSIONS

The Methodology proposed here is based on different classification tasks which are performed earlier. All Pairs Distance Calculation using Ordinal Classification is propsed here. All Pairs Distance Calculation checks all the possible pairs of combination to check whether randomly set data belongs to particular class or not. Here reformulation of data point is taken in to account, since proposed work rearranged the data based on minimum distance found, similarity vectors. Proposed methodology uses data set like HeartDisease, Thyroid, Contact-Lenses, Pasture, to validate the proposed method and it has been proved that it is superior when comparing with

other classifiers. Kernal Discriminant Analysis, Logistics regression, J48, CVR, MCC were applied to appraise this conclusion. From this, the superiority of the proposal for All Pairs Distance Calculation Using Ordinal Classification has been confirmed as the best classification method in the Process of Disease Prediction.

ACKNOWLEDGMENTS

Thanks to Sathyabama University for their support in Publishing this work. Also wewould like to express our great appreciation to all the members of our laboratory for their technical insight & simulating ideas, which is greatly contributed to the successof our research.

REFERENCES

1. BinGu, Victor S. Sheng, KengYeowTay, Walter Romano, and ShuoLi,"Incremental Support Vector Learning for Ordinal Regression",IEEE Transactions On Neuralnetworks, 2007; **6**(1).
2. WojciechKotBowski and Roman SBowi ´nski,"On Nonparametric Ordinal Classification with Monotonicity Constraints",IEEE Transactions On Knowledge And Data Engineering 2013.
3. B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink,"Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005; **27**(6): pp. 957–968.
4. Shereen Fouad and Peter Tino,"Ordinal-Based Metric Learning for Learning Using Privileged Information"IEEE Transactions On Knowledge And Data Engineering, 2013.
5. T. J. Hastie and R. J. Tibshirani, "Nonparametric regression and classification.Part II: Nonparametric classification," in From Statistics to Neural Networks: Theory and Pattern Recognition Applications (Computer and System Sciences), **136**, V. Cherkassky, J. H. Friedman, and H.Wechsler, Eds. New York: Springer-Verlag, 1996, pp. 70–82
6. T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multiclass classification by pairwise coupling," *J. Mach. Learn. Res.*, 2004; **5**,pp. 975–1005.
7. P. Tino and G. Dorffner:"Predicting the Future of Discrete Sequences from Fractal Representations of the Past," *Machine Learning*, 2001; **45**(2): pp. 187–217, Kluwer Academic Publishers
8. M.Perez-Ortiz,P.A. Gutierrez and C. Hervas-Martýnez,"Projection Based Ensemble Learning For Ordinal Regression ", *IEEE Transactions On Systems*, Man, And Cybernetics, Part B: Cybernetics,2013.
9. Paul R. Yarnold,"Maximum-Accuracy Multiple Regression Analysis: Influence of Registration on Overall Satisfaction Ratings of Emergency Room Patients"Optimal Data Analysis, LLC 2013, **2**: Release 1 (October 17, 2013), 72-75 2155-0182/10
10. Stefano Baccianella, Andrea Esuli and FabrizioSebastiani,"Evaluation Measures for Ordinal Regression",2009 Ninth International Conference on Intelligent Systems Design and Applications.
11. Yi Liu and Yuan F. Zheng,"One-Against-All Multi-Class SVM Classification Using Reliability Measures",2005
12. Kemal Polat and SalihGunes," A new feature selection method on classification of medical datasets: Kernel F-score feature selection", *Journal of Expert Systems with Applications*, 2009; **36**: PP.10367–10373.
13. W. Chu and S. S. Keerthi: "Support Vector Ordinal Regression," *Neural Computation*, 2007; **19**(3): pp.792–815.
14. JABEZ.J1, DR.ANADHA MALA.G.S2, " A study on Genetic-fuzzy based Automatic Intrusion Detection on Network Datasets " , in the proceedings of the International Conference on Software Engineering and Mobile Application Modelling and Development (ICSEMA – '12),pp 353-362, ISBN : 978-81-8424-812-8,IEE Explorer,IET Digital Library
15. Y. BevishJinila, K. Komathy, "Cluster oriented ID based multi-signature scheme for traffic congestion warning in Vehicular Ad hoc Networks ", *Advances in Intelligent Systems and Computing*, 2015; **338**: pp. 337 - 345.
16. S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures forordinal regression," in Proceedings of the Ninth International Conferenceon Intelligent Systems Design and Applications (ISDA 09), Pisa, Italy,2009
17. L.MaryGladence,K.Ravi,M.Karthi ,"Heart Disease Prediction using Naive Bayes Classifier –Sequential Pattern Mining " *International Journal of Applied Engineering Research (IJAER)-sep*,2014