Prediction and Classification of Patient Data Using Mutually-Correcting Nonparametric Model

Dr.Kasthuri
Assistant professor, department of computer applications, Bishop Heber College, Tiruchirappalli, India.

Sreejith.R
Research Scholar, Department of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India.

Abstract – Patient care systems are widely increased due to large population and huge health issues. The patient health records and the details should be managed properly fast quick data analysis and retrieval. The proposed system is aimed to apply machine learning techniques to effectively predict and manage patient flow. A new patient care framework is proposed with the machine learning techniques, which performs effective feature selection, prediction and formulates the estimation problem via non-linear model The proposed approach enhances the existing technique alternating direction method of multipliers (ADMM) by considering class imbalance issues. To achieve the above, a new framework named as non-parametric self sampling technique (NPSS), which consist of three algorithms such as OSCI-Over Sampling Class Imbalance with non-parametric model and auto regressive hidden markov model is proposed. This effectively mines the patient records and predicts the patient flow in various care units. The proposed work handles large, dynamic and patient sequential data with high prediction accuracy. In addition, a strong and hidden feature selection and learning is performed by applying the auto regressive techniques with hidden markov model. The class imbalance on the learning and prediction model is concentrated. This blends the training data and regenerates the dataset. Finally the proposed system shows the performance in terms of accuracy of predicting the destination critical unit care.

1. INTRODUCTION

The demand of care units are increased due to large population [1]. Effective patient care will reduce the waiting time in the emergency unit. This data collection and management is very crucial and need many improvements and enhancements. So, an effective patient care system is necessary. Data mining techniques are well adopted for many types of applications, specifically for the medical field. In the medical domain, predicting patient flow in the emergency unit and managing their clinical improvements and treatment progression along with the available emergency care resources are very important. There is no specific machine learning technique is available to handle such emergency care unit, so the proposed work initiates a data mining technique to predict patient flow in the healthcare domain. The proposed work aims to address the problem of patient flow prediction from the random sequential medical records. It effectively mines the continuous document clinic documents and predicts the patient flow and required resources.

2. PROBLEM DEFINITION

The main problem of patient flow prediction is handling the continuous patient health documents. Solving this issue will help to perform quick decisions and proper planning of hospital resources. However, solving these issues is very tough task due to many factors like huge data set, lack of resources, and uncertain document flows. And it also creates several issues on machine learning techniques such as time sensitivity, learning issues [2][3][4], classification and prediction issues. These techniques have several steps such as feature selection, data sparsity and class imbalance detection, and prediction techniques. The first feature selection is performed on the patient profile, treatments, medications and diagnosis reports. The correlation between these features is not completely defined, and among all the features, it is difficult to find the important factors that will detect the prediction process. Every feature selection process has generally depends on a specific model of patient feature selection process.

3. LITERATURE SURVEY

In the paper [5] authors have proposed an improved Kernel Learning with Individual Physique Indicators (KLIPI) model for mining medical record. This model is based on multi-dimensional Hawkes model. This captures the incidents of the patient diseases and their past medical history. The KLIPI model indicates the relationship between different kinds of diseases and disease evolvement as well. The authors have introduced a parameter representing individual physique and use Gaussian kernel density estimator for the estimation of the kernel function. The experiments of the KLIPI shoes the modifications both increase the accuracy of prediction on disease evolvement. The KLIPI Hawkes model outperforms Markov in disease prediction. However, there are several drawbacks and backlogs are found in the techniques. The approach is not suitable for continuous and time series data, and it creates class imbalance issues while prediction process is made. The accuracy of the prediction is very low.

In the paper [6], authors have concentrated on the class imbalance problem by adjusting the f-measure and kernel scaling. Here the author discussed various issues and impact of class imbalance. Authors have presented a classification procedure based on Support Vector Machine. The approach reduces the misclassification issues. However, the approach doesn't consider sequential datasets. And the multiple health features are not handled in this approach.

In the paper [7], authors have developed an ensemble method for dealing with the binary-class imbalance problems. Different from the conventional sampling methods, cost-sensitive learning methods, and Bagging and Boosting based ensemble methods, the method does not change the original class distribution, and does not suffer from information loss or unexpected mistakes that may be caused by these conventional methods via increasing the minority class instances or decreasing the majority ones. This technique, firstly converts the imbalanced binary-class data into multiple balanced binary-class data. This is achieved by applying random splitting or clustering to the majority class instances. After that, a specific classification algorithm is applied to the multiple balanced binary-class data to build multiple classifiers. Finally, the classification results of these binary classifiers for a new data are combined with a specific ensemble rule. Five new ensemble rules including MaxDistance, MinDistance, ProDistance, MajDistance and SumDistance, which depict the relationship between a new problem and the historical data, are presented.

In the paper [8], authors have used random balance technique, which is an ensemble approach with numerous existing classifiers to handle imbalance data's. The main idea behind this method is combining the features of different classifier and performs random selection over it. The new ensemble method named as RB-boost combines the existing Adaboost algorithm. This randomly selects data proportion for combining and generating new one. Authors have used many dataset and performed class imbalance data classification. The main drawback in the RB-Boost technique is it doesn't considered the intrinsic characters and noise elimination process were not used. These limitations need to be studied with dynamic electronic patient health records.

In the paper [9], authors have concentrated on patient flow detection from the electronic health records [EHR]. However, there are several techniques have proposed earlier to handle different types of machine learning algorithms, this paper is developed with NPSS to handle the patient health record, prediction of patient flow along with the resource prediction approaches. The authors have also concentrated on the maximum likelihood estimation process via discriminative learning algorithm. This provides an efficient learning method on HER. The approach obtains high accuracy in the given dataset samples and certain parameters have not yet performed on such factors like time duration.

The authors have used the *alternating direction method of multipliers* (ADMM) [10], which is an algorithm that resolves the convex optimization problems. The ADMM used Discriminative Learning of Mutually-Correcting (DLMC) algorithm with parametric approaches. This reduces the problem by segmenting them into small chunks. This process makes the application execution easier. From the detailed analysis, the proposed system is designed and developed NPSS approach. This approach aims to mine the EHR with dynamic flow records.

The existing mutually correlated process is replaced with several other approaches such as Modulated Poisson processes (MPP) [12], which applies the logistic regression in multinomial way. And another approach is Self-correcting process (SCP). Similar to the MPP method, the SCP method replaces the mutually-correcting process with the self-correcting process.

4. PROPOSED SYSTEM

Patient care systems are widely increased due to large population and huge health issues. The patient health records and the details should be managed properly fast quick data analysis and retrieval. The proposed system is aimed to apply machine learning techniques to effectively predict and manage patient flow. A new patient care framework is proposed with the machine learning techniques, which performs effective feature selection, prediction and formulates the estimation problem via non-linear model. The proposed approach enhances the existing technique alternating direction method of multipliers (ADMM) by considering class imbalance issues. The followings are the contribution of the proposed system.

- Auto-regressive model with hidden markov model, which improved and developed as a nonparametric model.
- Sampling technique is used to perform the class imbalance issues.
- Aim to improve the prediction accuracy.

The proposed system handles the dynamic event sequences by adopting new techniques. This has the ability to handle temporal dynamic events with high dimensional view. An improved discriminative learning based algorithm for the point process model is proposed, which combines the auto regressive and hidden markov model. The proposed work handles the feature selection problem by adopting the previous technique [11] and the imbalance of data are considered in the proposed learning algorithm. Finally, the sampling techniques have been proposed to handle the data imbalance problem, this effectively improves the accuracy of the prediction.

4.1 Auto-regressive model with hidden markov model:

The existing HMM (hidden Markov Model) is enhanced and developed as a nonparametric model with the help of auto

regressive techniques. This helps to find the unobserved or hidden states of the patient health record. This analysis is made with the historical information's of the patient such as medical diagnosis; treatment information etc. this process effectively summarizes the patient flow from the historical datasets. This can effectively find the probability of observed sequences. The improved Auto-regressive(AR) model with hidden markov model assumes the probability of the next state at the time series "i" is depends on the current hidden state and high priority historical information shown in equation 1.0. For all time series, the hidden factor is analyzed and prioritized by the AR- values.

$$H(Mi+1|Mi, Mi-1,\ldots M0)=H(Mi+1|Mi) \quad Eq(1.0)$$

This also represents the statistical correlation between two states at the time of analysis. The hidden state is denoted as "S", and the emission are denoted as "E" for the time series "i". This will facilitate the dependencies between every state Si….Sn with the Ei…En; using this combination and correlation process, the system avoids the over fitting problems. The high priority features are integrated in to the prediction process, so the proposed system will reduce the prediction delay.

4.2 Proposed sampling technique:

With the help of the auto regressive hidden markov model, the hidden features are identified and thus improve the prediction model. The class imbalance issues have discussed by many authors and provided many techniques on it. The proposed system utilizes a new sampling technique, which adopted with the hidden feature detection process. This contains the following steps.

Step 1: Class Division:

The data sets are partitioned into majority and minority subsets. As the concentration is on over sampling, the minority data subset was taken for further analysis and data increasing.

Step 2: Hidden Feature Subset Selection

In the next step of the proposed work finds the noisy data from the majority subset and that will be treated first. The minority class analysis is performed with the basic comparison and effectively divides the total dataset into equal parts. . One of the ways for finding the weak instances is to find most influencing attributes or features and then removing ranges of the noisy or weak attributes relating to that feature. The most influencing attribute can be found is by using the existing feature selection techniques.

Step 3: Selection of Feature Class label:

The noisy records relating to those least scored features can be chosen from the data set and this can be eliminated to reduce the majority class samples.

Step 4: Developing the Modified Data Set and classification:

The minority subset and the stronger majority subset are combined to form a strong and balance data set, which is used for learning a base algorithm. In this case C5.0 was used as the classification algorithm. The C5 algorithm avoids the over fitting problem and reduces much iterations. The proposed system utilized the non-parametric machine learning approach along with the AR-HMM.

The algorithm 1: The proposed algorithm presents all the above said steps for implementation of patient flow prediction and classification framework. The algorithm is given below,

Algorithm 1: OSCI

Input: A set of training samples, which contains Class C, and total instances I in every class.

Output: sampling and classified result

Steps:

1. Initiate data collection D
2. Find the class imbalance ratio IR and apply feature selection (Fs) on subset S.
3. Find the correlation between the features F.
4. Define threshold T for IR ratio ,
5. If (Ci<T)
6. Perform oversampling.
7. Repeat step 5 for every class
8. Train and learn a base classifier C5
9. Return classified result R.

The class imbalance issue is handled by the OSCI algorithm above. The proposed system performs the prediction and classification of patient flow based on the mutually correcting process. For example a patient in a specific care unit has the highest probability to be transferred to another care unit. This considers the previous time duration and predicts the expected move can be performed by the mutually correcting process. So it concludes the correlation between various features and classes should be identified. Using the AR based HMM, the hidden features can be detected. Finally a non- parametric machine learning algorithm has used to classify the patient data and helps in flow prediction.

5. EXPERIMENTAL RESULTS

Dataset: the implementation of the OSCI is performed on different types of patient health care dataset. The major dataset "Post-Operative Patient Data Set" is collected from UCI repository [13] and then modified to show the class imbalance problem. The dataset consist of 8 attributes, 90 instances and 3

classes. The dataset has unequally distributed. The class I contains 2 attributes, class S contains 24 attributes and the class A contains 64 attributes. By using this dataset, the class imbalance issue on patient classification is performed.

The proposed system is compared with the existing technique ADMM with DLMC algorithm. For proving the performance of the proposed system along the existing method, the following measurements have measured:

Prediction accuracy: The prediction accuracy P for each dataset d and the overall accuracy Pd are calculated as

Pd = records correctly predicted (CR)/ total records (n) Eq(2.0)

From the above formula, the total accuracy is calculated and compared.

Table 1.0 accuracy comparison table

Technique	Accuracy (%)
MPP	0.656
SCP	0.618
DMCP	0.652
NPSS(OS-AR-HMM)	0.694

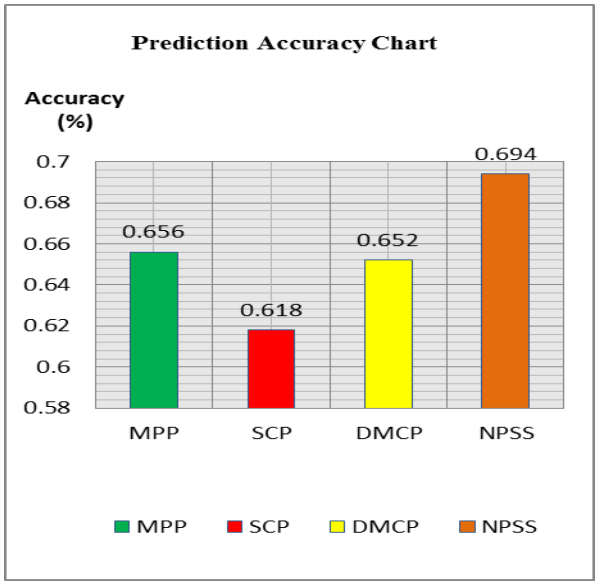

Figure 1.0 Overall prediction accuracy for various methods

The overall performance of the proposed system NPSS is compared with the existing MPP, SCP, DMCP. The proposed system shows high accuracy than the existing techniques.

6. CONCLUSION

Classifying patient electronic record and prediction patient flow in health care unit is done in the proposed system. The proposed system provides a non-parametric approach for predicting and classifying patient record. The approach utilizes the auto regressive hidden Markov model and over sampling technique to reduce the issues of class imbalance. Finding hidden features and performing appropriate oversampling is performed in the proposed system. Even though, there is several techniques have used in the literature to perform the event sequence analysis, the proposed AR-HMM performs better accuracy with the non-parametric machine learning algorithm C5.0. The experimental results on the accuracy metric are described with synthetic patient records.

In future, the approach can be elaborated with real time patient care datasets along with the other classification algorithms. The proposed over sampling technique can be replaced by some other class imbalance techniques and that can be compared with the proposed system.

REFERENCES

[1] Dharshi, Anisha. "The future of emergency care in the United States health system." *Journal of pediatric surgery* 41, no. 11 (2006): 1793-1798.

[2] Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets." *ACM Sigkdd Explorations Newsletter* 6, no. 1 (2004): 1-6.

[3] Kumar, B. Senthil, and R. Gunavathi. "Comparative and Analysis of Classification Problems." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7, no. 8 (2017).

[4] K. Sethi and G. Sarvarayudu, "Hierarchical classifier design using mutual information," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 4, pp. 441–445, 1982.

[5] Zhao, Yinan, Xiaojuan Qi, Zhengzhe Liu, Ya Zhang, and Tao Zheng. "Mining medical records with a klipi multi-dimensional hawkes model." In *KDD 2014 Workshop on Health Informatics*. 2015.

[6] Maratea, Antonio, Alfredo Petrosino, and Mario Manzo. "Adjusted F-measure and kernel scaling for imbalanced data learning." *Information Sciences* 257 (2014): 331-341.

[7] Sun, Zhongbin, Qinbao Song, Xiaoyan Zhu, Heli Sun, Baowen Xu, and Yuming Zhou. "A novel ensemble method for classifying imbalanced data." *Pattern Recognition* 48, no. 5 (2015): 1623-1637.

[8] Díez-Pastor, José F., Juan J. Rodríguez, César García-Osorio, and Ludmila I. Kuncheva. "Random balance: ensembles of variable priors classifiers for imbalanced data." *Knowledge-Based Systems* 85 (2015): 96-111.

[9] Xu, Hongteng, Weichang Wu, Shamim Nemati, and Hongyuan Zha. "Patient flow prediction via discriminative learning of mutually-correcting processes." *IEEE transactions on Knowledge and Data Engineering* 29, no. 1 (2017): 157-171.

[10] Wang, Yichen, Robert Chen, Joydeep Ghosh, Joshua C. Denny, Abel Kho, You Chen, Bradley A. Malin, and Jimeng Sun. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1265-1274. ACM, 2015.

[11] B.Senthil Kumar and Sreejith.R. "A Novel Machine Learning Approach to Diagnose Type 2 Diabetes and Different Clinical Datasets Using Weighted Genetic PCA Methods." *IJIRCCE , Volume 4*, issue 11 (2016): 19776-19782.

[12] B. F. Cole, M. Bonetti, A. M. Zaslavsky, and R. D. Gelber, "A multistate markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness," Statistics in Medicine, vol. 24, no. 15, pp. 2317–2334, 2005.

[13] https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient