www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 10 Oct. 2016, Page No. 18697-18701

A New Hybrid Prediction Approach For Enhance Prediction Accuracy of Complex Data

Akash Sharma[1], Nikita Jain[2]
[1]M.Tech Scholar, Department of Computer Science & Engineering, G.I.T., Jaipur, Rajasthan, India
[2]Assistant Professor, Department of Computer Science & Engineering, G.I.T., Jaipur, Rajasthan, India

Abstract: ***Over past few decades a lot of approaches have introduced for data prediction practice and many of researchers are continuously introducing their unique ideas on a daily basis for enhancing the power of prediction system in different areas but each and every approach has its own limitation. One of the most limitations of existing prediction approaches is that they are designed with the strength of a single classification technique and are not enough to handle different types of data efficiently thus the cost of system is high and field still lacks with high rate of prediction accuracy. This dilemma of existing prediction systems has consider into this paper with proposing an effective and efficient data prediction system that enhances data prediction accuracy in a significant way.***

Keywords: Data mining, Prediction Techniques, Decision Tree, Classifier, NaiveBayes.

1. Introduction

In modern time with the rapid prologue of new expertise the use of digital systems and its applications have attain a huge growth in almost each and every field of the work. A regular growing amount of peoples use these systems as a source of information and there is like to be impractical for an individual and organizations to accomplish their daily tasks without relying on the conveniences provided by these systems. Although still users have doubts about truthfulness of the proliferate contents. In some crucial working fields such as in the area of medical where incorrect or deferred predicted information may cause of danger or more harmful to a person, therefore quickly and accurately prediction of syndrome signs in time is must for shaping the trust. On the other hand, due to various causes the software appears with many defects. Predicting defects of an application at an early stage minimize the cost, time and pick up the overall effectiveness, attain a huge growth in recent days [1]. Additionally precious information is always prone over the network. Therefore quickly and accurately finding of the content that influences the trust is very important. In direction to enhance the prediction accuracy of complex data an endeavor has been made in this dissertation. The proposed mechanism ensembles the functionality of naïve bayes and KStar in a layer forms to test data for its correct prediction.

2. Prediction System & Techniques

Technically, the term prediction is based on the accuracy of verdict connection or patterns among dozens of fields in large databases. For such type of achievement the regular process of database updating and machine training are essential key features. Usually, for a machine learning process, two foremost components, presented in figure 1 are most popular routines [4].

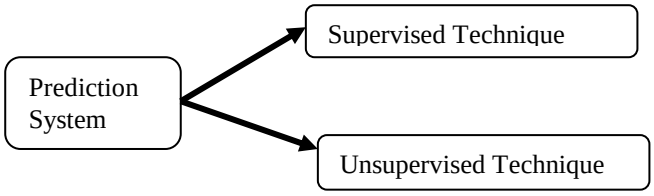

Figure 1 Data Prediction Techniques

Usually Supervised learning technique is a machine learning (Data Mining) task from supervised training data. The technique execute in a regulation to generalize from the training data to correctly determine the class labels for unseen instances. It offer training example where each example is a pair, consisting of an input object and a desired output value also called the supervisory signal. The algorithm examines grounding statistics and constructs an inferred function, which is called a classifier if the output is discrete or a regression function if the output is continuous [2].

Unlike supervise learning process the unsupervised learning technique annoying to discover concealed constitution within unlabeled data. Figure 2 present the dissimilarity between supervised and the unsupervised technique.

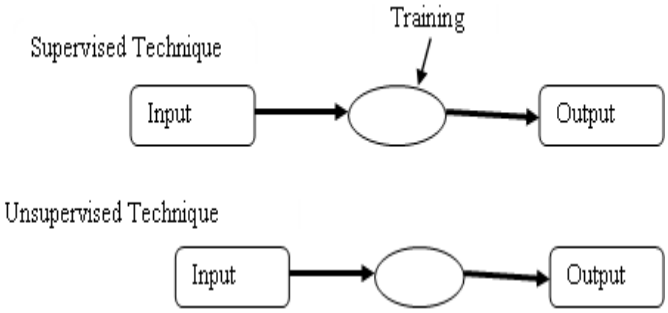

Figure 2 Supervised & Unsupervised Techniques

3. Obtainable Data Classification Algorithms

A number of researchers have proposed huge amount of classification algorithms, which can be divided into a broad range of categories. Some of the accessible classification algorithms are presented in figure 3.

DOI: 10.18535/ijecs/v5i10.60

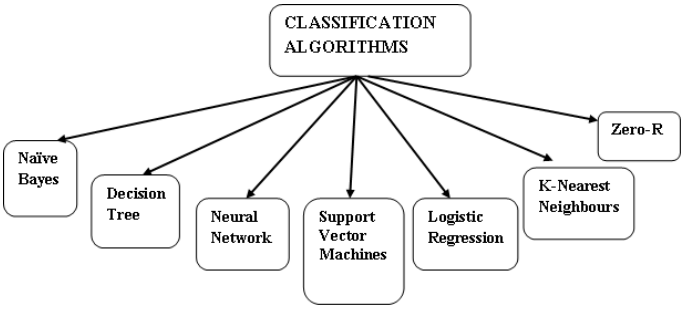

Figure 3 Classification Algorithms

3.1 Naïve Bayes

A Naïve Bayes algorithm is one of the most popular and successful supervise learning as well as a statistical classification method of data mining. This approach is simply based on the probability theory, employ Bayesian theorem with burly self-governing assumption for learning to classify text documents [3]. Due to its simplicity, technique can be build straightforwardly and most suitable in a high inputs environment for calculating the probability by using following equation.

$$P(z|v) = \frac{P(v|z)P(z)}{P(v)} \quad (2.1)$$

i.e.

$$P(z|v) = P(v_1|z) \times P(v_2|z) \times \ldots \times P(v_n|z) \times P(z) \quad (2.2)$$

Where :

P(z) and P(v) represent the prior probability of class and predictor. At the training time of model probability of each class is measured by finding total occurrence of it in training dataset, known as "prior probability". The P(z|v) shows the posterior probability of class z (target) given predictor v (attribute).

3.2 Decision Tree

A decision tree is another one popular data mining technique which inspects data and makes predictions. For concluding a course of action or present statistical probability this technique recursively split a dilemma into subsets until it solve problem directly without any trouble. The attributes of a problem becomes as nodes of constructed tree and their comparative values determine the paths of the tree. Due to its simple working functionalities and healthy outputs this technique has lead over many techniques.

3.3 Neural Network

In this technique a large number of highly interconnected processing elements (neurons) form a network and work together for accomplishing a task. Due to high learning efficiency the neural network technique produce effective results in comparison with the other classifier algorithms when the majority of variables are weakly relevant. Several of neural network frameworks can be categories as Feed-forward & Recurrent neural networks. The Feed-forward neural network is a simple network where information travels in one direction only, from input to the out neuron by using intermediate neurons. On the other hand in recurrent neural networks established connections between the units shape a directed cycle.

3.4 Support Vector Machines

The support vector machine (SVM) [4] is a training algorithm for learning classification and regression rules from data. This technique was introduced in the year 1960s for classification. It employs statistical learning theory to classify all training instances correctly by separating them into correct classes [5]. This technique supports both regression and classification tasks and performs the work on base of the structural risk minimization principle, closely related to regularization theory.

3.5 Logistic Regression

This technique is a powerful modeling tool, considered as a simple standard statistical approach which uses one or several predictors (numerical and categorical) for prediction. For predicating the value of a binary variable the technique of linear regression is not an appropriate solution. This technique has been applied in many data mining classification and prediction problems where the dependent variable (target) has just two values, such as: 0, 1 or in form of Yes & No or True/False.

3.6 Nearest Neighbors - Classification

This classification technique is a non parametric lazy learning algorithm, does not require any training phase and not makes any assumptions for generalization. It is a simple algorithm which stores all available cases and classifies new cases based on a similarity measure. Typically in this approach on the base of distance metric new arrival instance is compared with the stored instances for assigning class, arrival instance join class by its closest instance. Sometimes more than one nearest neighbor is used, and the majority class of the closest k neighbours (or the distance weighted average, if the class is numeric) is assigned to the new instance. However, on basic recognition problems this technique perform healthy but not suitable for real time application, may perform leisurely with huge training examples. Apart from this, technique does not learn anything from the training information and are not robust to noisy data, which can result in the algorithm not generalizing well.

3.7 Zero-R

For the prediction of statistics the mechanism of Zero-R classification relies on its training set. Typically, this classification method has not power of prediction, simply useful as a benchmark to compare the performances of other classifiers. It simply predicts the majority category (class). With the numeric statics the technique predicts the average value of the target attribute from the training set.

4. Related Work

Over past few decades, the speedy uses of digital system in different areas of work have increases the amount of data. In real time these systems are the most source of information and play an important role in decision making. To accomplish this task a more accurate prediction and huge collection of information is always required. To analyses the prediction

power of existing classification algorithms an investigation has compare performance of Naïve Bayes and Rule Induction technique [6]. The outcomes have presents the comparative qualities of existing algorithm in term of predicting software components defects. For the analysis part the authors have used 10 separate data sets. In same direction another investigation carried out a number of simulations to show an efficiency of ANN technique in the field of a prediction system design [7]. However this proposed technique enhances the performance of accessible system but produce huge amount of false alarms which is a major limitation of proposed approach.

A good number of researchers have use different approach for improving the quality of rule based prediction system. Typically the rule based techniques take out only important features set from a variety of audit streams and use these selected features for building an perform prediction. In proposed approach the authors has apply different methods such as clustering, classification to analyze the data but still approaches face some limitation such as clustering technique has limitation that it cannot be easily used with symbol features, the observation must be numeric. It considers the features independently and unable to capture the relationship between different features of a single record, which degrades attack detection accuracy [8],[9].

For gaining an accurate and high prediction rate another two separate approaches have use multi-classification technique [10],[11]. Apart of combining the functionalities of existing classification mechanism the authors of proposed approaches have incorporates a feature reduction method. They present separate ways for the selection of suitable feature set from the whole set of an attributes. They have performed number of simulations with different parameters and compare the outcomes of their proposed approaches with the recital of existing techniques, decision tree, naïve bayes, neural network technique. The comparative outcomes have presented the importance of proposed algorithms in direction to improve prediction accuracy of data with generation of fewer false alarms However, approaches enhance the rate of prediction but fail to making quick decision, require more time for an simulation work. Apart to requiring more time for prediction process the recital of presented approach has not compare with the new dataset, for simulation work authors have use same dataset as they use at the time of machine training.

A group of authors has demonstrated a Rainfall Prediction system [12]. For system training the authors of this investigation has customized neural network and given a name Data Core Based Fuzzy Min Max Neural Network (DCFMNN). After the modification they have use this dataset for pattern classification. According to authors the traditional neural network called fuzzy min max neural network (FMNN) creates hyperboxes for classification and predication which have a dilemma of overlapping neurons that resoled in DCFMNN to give greater accuracy. For removing such issues of classical approach the authors have composed forming of hyperboxes. They use two kinds of neurons called as Overlapping Neurons and Classifying neurons, and classification used for prediction. For each kind of hyperbox its data core and geometric center of data is calculated. The simulated results illustrates that presented technique has produce high accuracy and strong robustness.

Many of investigators has also demonstrates a number of techniques with their reward and limitation [13]. Apart of presenting of an overview of traditional algorithms much of authors have demonstrated comparative analysis of several of techniques [14]. They illustrate different techniques in detail with their working methodologies and have also discussed the limitation of presented methods. In same context a new investigation has present a survey on traditional prediction approaches with defining inadequacy of such technique in a new way which may be helpful for new investigators to investigate new approach with high accuracy and reducing rates of false alarms[15].

5. Dare for Accessible Classification Algorithms

However, a number of prediction approaches have been proposed by the numerous researchers and still unique ideas has comes on a daily basis for enhancing the power of existing mechanism in different areas but each and every approach has its own limitation. One of the most issues associated with the on-hand prediction techniques is that most of them are designed with the power of single classification algorithm. Therefore the proposed mechanisms faces same issue as associated with the existing classification techniques such as not suitable for handle different data types or some are fail to response in real time scenario. Apart from this numerous techniques associate different significant limitation which can be express as

- Most of techniques are not suitable for an environment of high inputs.
- Required manual effort on a regular basis for updating its functionalities.
- Produce low accurate results and are inflexible.
- Number of approach can detect the signs which were used in train dataset.

Additionally designing cost of the most of on hand prediction system is high and field still lacks with high rate of prediction accuracy. However, a number of researchers applied numerous techniques to enhance the accuracy of existing prediction system with reducing its faultiness but different classification algorithms perform differently and not so much are useful to give sign about the actual number of accurate prediction. This dilemma of existing systems consider in this dissertation work by offering an effective and efficient prediction classification mechanism which enhances the prediction accuracy with generating the low false alarms.

6. Proposed Prediction Technique

In context to enhance the prediction rate with generating low false alarms proposed algorithm has combine functionality of two popular classification techniques Naïve bayes algorithm and KStar and are designed in a layer format. Each layer is designed with a separate classification technique and trained with only useful subset of feature instead to use a complete feature set at each and every layer. The subset features are

DOI: 10.18535/ijecs/v5i10.60

selected by using the existing feature selecting technique known as CFS (Correlation-based Feature Selection) with the BFS (Best First Search) method. After completion the feature subset selection process the selected features are applied with the naïve bayes classification algorithm for prediction of desired facts. Once the process of naïve bayes algorithm has been completed then correctly predicted information has been detached from the supplied data set and remain statistics is used with the same feature set at the next layer that is designed with a different prediction technique named as KStar for enhancing the rate of accurate prediction. In proposed approach if any fact is not accurately predict at a first stage of inspection the next layer due to alter the classification technique increases the chance of correctly prediction.

7. Assessment of Results

For the evaluation of effectiveness of designed approach over the other accessible approaches a number of different experiments have been carried out and outcomes has compared with some of selected algorithm against designed approach on the base of confusion matrices. The predication performance of the proposed scheme is calculated on the base of Accuracy, measured by the following equation

$$\text{Accuracy} = \frac{\text{TruePossitive}+\text{TrueNegative}}{\text{TruePossitive} + \text{TrueNegative} + \text{FalsePossitive} + \text{FalseNegative}} \quad (4.1)$$

Where, TruePossitive, TrueNegative, FalsePossitive, and FalseNegative refers to the actual percentage that were predicted by the classification model. A classification algorithm may wrongly classify some instances because of the bias nature of the dataset but may appear to be accurate. As a result to compare an accuracy TruePossitive, TrueNegative, FalsePossitive, and FalseNegative rates has been calculated to find out the main accuracy of simulated classification algorithm.

Table 1 Relative Performance over Breast-Cancer Dataset

S.No.	Prediction Techniques	Accuracy (%)
	Proposed Approach	95
	Naïve Bayes	72.38
	Decision Tree (J48)	73.08
	IBK	71.68
	One R	65.73
	KStar	73.78
	Hoffeding Tree	71.33

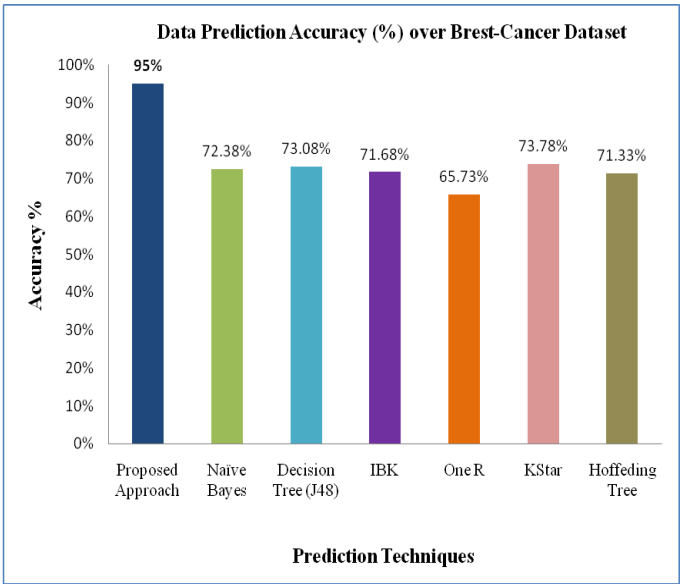

Figure 4 Prediction Recital of Designed Approach

In order to evaluate the efficiency of designed approach in real time with different parameters another experiment has done with using the different dataset i.e. kidney disease. The figure 5 has demonstrates a graphical illustration of the statistics of above table 4.8 to explain efficiency of designed approach in better way.

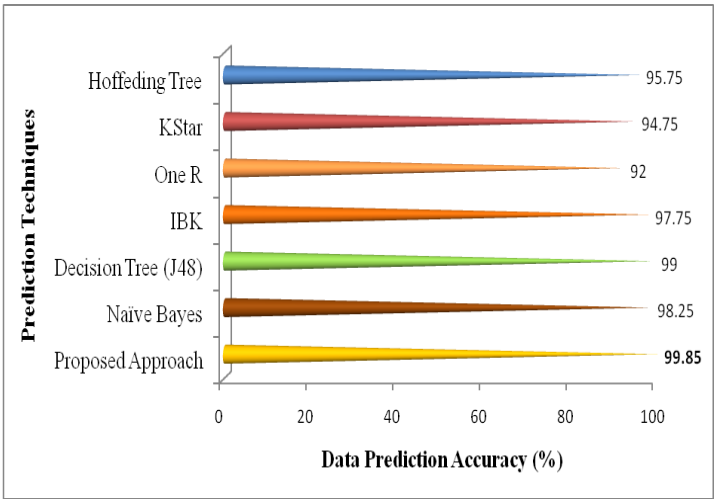

Figure 5 Prediction Performance of Proposed Approach over Kidney Disease Dataset

8. Conclusion

The results that are demonstrated in above table and figures have explained that designed approach of this investigation has improved data prediction accuracy rate significantly. All the results have been evaluates with the same parameters, means each prediction technique has been evaluates with the same dataset. Apart of using same dataset the other parameter of evaluation such as use of 10 fold cross validation technique without training the introductory system remains same to evaluate the performance of each data prediction algorithm. However all the used dataset for the evaluation purpose have different property such as different types of attributes and the instances. Apart of these some of the dataset have different nature, have the distinguished field data but all that have not been affected the performance of the proposed approach.

References

[1] Daniel Dawson, Nathan Hawes, Christian Hoermann, Nathan Keynes, and Cristina Cifuentes. "Finding bugs in open source kernels using parfait" Sun Microsystems, November 2009.

[2] T. J. Ostrand, E. J. Weyuker, and R. M. Bell. "Predicting the location and number of faults in large software systems", IEEE Transactions on Software Engineering, 31(4):340–355, 2005.

[3] Hetal Doshi, Maruti Zalte "Performance of Naïve Bayes Classifier Multinomial Model on Different Categories of Documents" National Conference on Emerging Trends in Computer Science and Information Technology (ETCSIT) 2011 Proceedings published in International Journal of Computer Applications® (IJCA).

[4] O. Chapelle, P. Haffner and V. N. Vapnik, "Support vector machines for histogram-based image classification", IEEE Transactions on Neural Networks, vol.10, no.5, (1999), pp.1055-1064.

[5] Perlich, C. Provost, F. Simonoff. J S "Tree Induction vrs. Logistic Regression: A learning-Curve Analysis". Journal of Machine Learning Research. 2004. Vol. 4. p.211 – 255.

[6] Tim Menzies, Jeremy Greenwald, and Art Frank. "Data mining static code attributes to learn defect predictors". Software Engineering, IEEE Transactions on, 33(1):2{13, 2007.

[7] Iker Gondra. Applying machine learning to software fault-proneness prediction. Journal of Systems and Software, 81(2):186{195, 2008.

[8] Animesh Patcha and Jung-Min Park. "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends", Computer Networks, 51(12):3448– 3470, 2007.

[9] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion Detection with Unlabeled Data Using Clustering", Proc. ACM Workshop Data Mining Applied to Security (DMSA), 2001.

[10] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu Ranga Vital "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms", International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013.

[11] S.Kharya, D. Dubey, and S. Soni "Predictive Machine Learning Techniques for Breast Cancer Detection", IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023 - 1028.

[12] Rajendra Palange, Nishikant Pachpute, Snehal Pandharbale, Shubham Gupta, Dr. Mrs. Swati Shinde "Rainfall Prediction using Data-Core Based Fuzzy Min-Max Neural Network for Classification" Rajendra Palange et al. Int. Journal of Engineering Research and Applications, Vol. 5, Issue 5, (Part -4) May 2015, pp.38-41

[13] Chaitali Vaghela, Nikita Bhatt, Darshana Mistry, "A Survey on Various Classification Techniques for Clinical Decision Support System" International Journal of Computer Applications, Volume 116 – No. 23, April 2015.

[14] Dr. S. Vijayarani, Mr.S.Dhayanand, "Data Mining Classification Algorithms For Kidney Disease Prediction" International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.

[15] Shubpreet Kaur and Dr. R.K.Bawa "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System" International Journal of Energy, Information and Communications Vol.6, Issue 4, 2015, pp.17-34

BIOGRAPHIES

Akash Sharma currently pursuing M.Tech (CSE) from GIT College Jaipur affilated to Rajasthan Technical University, Kota. He did B.TECH in Computer Science and Engg. from BMIT, Jaipur in 2010. His interested research areas are Data Mining. Computer Networks.

Ms. Nikita Jain obtained B.Tech. Degree in Computer Science and Engineering from UPTech University lucknow in 2005 and also completed M.Tech. in 2014 with same subject from Rajasthan Technical University, Kota. She has published Several research papers in reputed conference & journals.