Credit Card Risk Evaluation Using Incremental Learning Technique

Pallavi Kulkarni[1],Roshani Ade[2]

[1]Department of computer engineering,Dr. D. Y. Patil School of Engineering & Technology, Pune,
kulkarnipallavi4@gmail.com
[1]Department of computer engineering, Dr. D. Y. Patil School of Engineering & Technology, Pune,
rosh513@gmail.com

Abstract—Credit card is the popular mode of payment in financial field. With the increasing number of users' worldwide, risks on use of credit card has also been increased, where there is danger of stealing of credit card details and committing threats. Incremental techniques are desirable in recent machine learning areas like financial problems such as credit card risk assessment since amount of data and information is intensifying over the time. Scale up in learning can be achieved by updating classifier as and when training data becomes available. A smart technique known as ensemble technique has become popular, in which multiple classifiers are united in such a way that correct decisions are amplified and incorrect ones are discarded. Major focus of ensemble based techniques is diversity of classifiers that leads to reduction in misclassification. This paper presents ensemble based technique named as KIN algorithm, which handles credit data efficiently and finally distinguishes the bad customers from faithful ones in more accurate way.

Keywords-Incremental learning, Ensemble Technique, Credit threat evaluation

I. INTRODUCTION

The enterprise credit threat evaluation has long been contemplated as essential issue in the academic and business community. It has paved the path and became important factor for financial institutions to evaluate threats in credit, enhance cash flow, lower risks and take decisions at management level. To gain profit in case of financial institutes accurate results should be obtained and that is crucial in enterprise credit risk assessment. There are several data mining techniques that can be used for analyzing problems in real world applications like credit card fraud detection and discovering their solution in a scientific manner. If learning from enormous data which is progressing by sequential steps is desired, then incremental learning or online algorithms are best suitable and preferred. Learning from stream data needs a classifier that can be incrementally modified to get full benefit from newly arrived data, while simultaneously preserving performance of the classifier on matured data. Ensemble technique (multiple classifier system) is widespread in the area of machine learning, specifically in the incremental and non-stationary environment. Ensemble techniques need a mechanism which intelligently integrates results from multiple classifiers and produces a final decision as an output. This mechanism is known as voting technique and there are various ways to do this. Next sections in this paper cover details of incremental learning, ensemble technique, voting methods with respect to credit card fraud detection [1,7].

II. BACKGROUND

Several statistical methods are present that builds the credit threat assessment model. Supervised

learning techniques use instances, which have already been pre-classified in some manner. Classification is a supervised data mining technique that maps, data into groups of classes established in advance. Data evolves over time, as time goes on and rapid increase in volume of data is observed. So for such applications, incremental learning approach is desired. Upcoming subsections entails overview of various ways that are used for credit risk assessment including incremental learning technique.

2.1 Online Learning

If learning from enormous data which is progressing sequentially is required, then incremental learning or online algorithms are best suitable. Incremental learning algorithm is always fed with input data as the data arrives in sequence, the algorithm computes hypothesis by taking into account order of data arrival, each one gives details of all the data seen so far[2]. Note that the algorithm should be depend on prior hypotheses and the ongoing training data. Incremental learning algorithm should possess all the properties mentioned below:

1. The algorithm should gain knowledge which is additional in recent data
2. To train already present classifier, the algorithm need not access to the initial data
3. It should keep up the knowledge which it has previously learned (No catastrophic forgetting should occur)
4. New data may bring in concept class

This definition points out that, learning from stream data needs a classifier that can be incrementally modified to get full benefit from newly arrived data, while simultaneously preserving performance of the classifier on matured data underlining the fact to of stability-plasticity dilemma which describes how a learning system can be designed to stay stable and unchanged to immaterial affairs, while plastic (i.e. be able to change when necessary in order to deal with different situations) to recent and crucial data[9-10].

2.2 Multiple Classifier System

Incremental learning properties can be achieved by building ensembles of batch learners, where fresh batch learners can be learned on the fresh data, and then united through a voting system. Ensemble technique (multiple classifier system) is widespread in the area of machine learning, specifically in the incremental environment. An ensemble technique is extracted by merging diverse classifiers. There are various differentiating parameters who help to achieve diversity that in turn entitles each classifier to produce several decision boundaries. Appropriate diversity allows to gain different errors to be made by individual classifier and finally strategic integration of them can cut off the total error in the entire system.

The diversity need can be fulfilled by applying various approaches such as:

1. Training each classifier using several data chunks
2. Training each classifier using several parameters of a given classifier architecture
3. Training each classifier using several classifier models
4. Random Subspace method (training each classifier using several subsets of characteristics

The Ensemble can be build up in several ways like:

1. Bagging
2. Boosting
3. Stacked generalization
4. Mixture of experts

Ensemble approach can be used to address every challenge associated with incremental learning as a basic design model [3,11].

2.3 Voting Strategies

Fundamental issue in ensemble technique is a selection of appropriate rule to unite decisions from multiple experts. Voting rule is applied at the final step of ensemble system [8]. In literature, various voting rules are presented. Most commonly used examples are:

1. Geometric average rule (GA rule)
2. Arithmetic average rule (AA rule)
3. Weighted average rule (Weighted AA rule)
4. Median value rule (MV rule)
5. Majority voting rule (MajV rule)
6. Weighted majority voting rule (Weighted MajV rule)
7. Borda count rule (BC rule)
8. Max and min rule
9. SSC rule

III. PROPOSED FRAMEWORK

It is obviously necessary to pick optimal group of classifiers for ensemble based system and the selected classifiers should be diverse enough so that the classification mistakes of one classifier should not be repeated by another one. This is vital for multiple classifier system to take full benefit of the system's structure. When multiple decisions are united using voting technique, one expect to obtain accurate results based on the assumption that most of the experts are more probable to be right in their classification result when they go along with each other in their judgment.

Three incremental learning algorithms are used as base learners in the proposed system. They all use nearest neighbor classification technique which is pretty popular lazy learning based on learning by analogy. Lazy learning algorithms save the training instance and stay idle until it gets a test instance. As soon as the testing instance is stipulated, it computes generalization based on its similarity to the training instance [5]. These technique perform less work in training instance is provided and practice more work in testing duration. Such approach is also known as "instance based learning". The major advantage of lazy learning technique is that they are capable to learn instantly form tiny dataset. Additionally, they can solve multiple issues simultaneously and handle modifications in the problem arena. They are described in detail as:

3.1. K-Star

KStar is an instance based learning algorithm with some special feature added to it. KStar is different from other lazy learners because it uses entropy based distance measure to find nearest neighbor. This technique is inspired from information theory. Basic idea behind it is that distance between instances can be computed by the complexity of mapping one instance into other. There are two phases for such transformation. In the first phase, one instance is mapped into other instance and in the second phase, finite transformation is performed. An interesting functionality about KStar is that it is capable to tackle the problem of missing values which is common for many datasets. Generally for such problem, dataset need some sort of preprocessing. Kstar can deal with such datasets which is having missing values for multiple attributes. KStar handles this issue by presuming that missing values can be served as if they were, randomly chosen among the samples in the dataset. In other words, this approach uses the method of summing probabilities over every possible path. As a result, the algorithm resolves the smoothness issue and hence achieves better quality in the performance [4].

3.2. IBK

IBK/IB1 algorithm uses k-NN method for the implementation. This method exploits normalized distance measure for every attribute so that attributes on different scales have the similar influence on the distance function. Resulting distances may have ties. In that case the algorithm may return with more than k neighbors. Finally, the neighbors are voted for an ultimate result of classification.
To compute nearest neighbor, IB1/IBK goes for Euclidian distance metric. Only difference between IB1 and IBK is the value of last parameter, i.e. 'k'. K value is 1 in case of IB1 and it is more than 1 in case of IBK. K value is nothing but number of nearest neighbors the algorithm has to look for. This k value can be determined either by instinctively using leave-one –out-cross validation focus to an upper limit stipulated by the stated value or can be stated frankly in the object editor. Several search techniques can be exploited to improve the speed to look for nearest neighbors.

3.3. NNGe

NNGe stands for Non-Nested Generalised Exemplars. Generalised exemplars are nothing but a restricted class of
samples that share the similar concept and are close in proximity in the feature space.
This algorithm has its focus on uniting the samples together, which results in reduction in their frequency. It does not store the instances as it is, rather finds generalized tuples to integrate them.
New instances can be handles in one of the two ways:
1. Add them into already present generalized instances
2. If they are covered by generalized instances already, then discard them, repeating of examples is not allowed.
NNGe first classifies instances and then generalizes them. It is obvious that generalization is carried out by NNGe each time a new instance is added to the database, by combining it to its nearest neighbor of the same class.
Generalization methods include nesting and overlap technique. NNGe does this without nesting or overlap, instead it uses a method which is based on formation of hyper rectangles in sample space that represent conjunctive rules with internal disjunction. In other words, NNGe implements bounding group of examples using this hyperrectangles[6].
The proposed system (KIN) initiates by forming a group of three classifiers (NNGe, IB1/IBK, KStar). When data tuple arrives, the framework forwards it to and accept prediction form each expert in the ensemble. As this is incremental learning framework, the system continually updates its decisions, it continuously gets a pattern, computes its value by considering majority vote of the classifier's hypothesis and finally computes final outcome accordingly. The KIN system of ensembles is depicted schematically in fig. 1, where *H** is final classification decision after computation of majority voting, *hj* is the classification decision of individual learners in the ensemble, x is the sample used for training the system, and y* is the final class label produced as prediction by the proposed system (KIN). A feature of this ensemble based system (KIN) that attracts to non-data mining people is , this system has less number of runtime parameters that need to be tuned. Hence the algorithmic system (KIN) is easy to use. Additionally, the proposed system (KIN) supports parallelism and distributed computing. The intuitive is to use one learner per machine for extensive use.
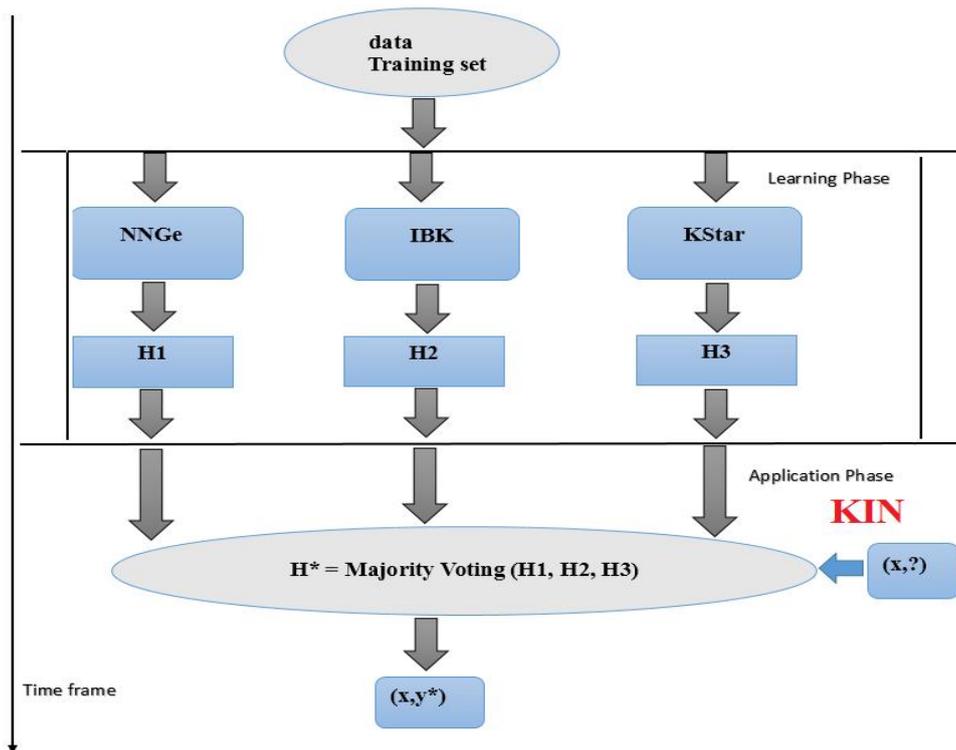
Fig 1.Architecture of the Proposed Framework (KIN)

The support for parallel computing and distributed environment leads to improved speed and allows the system to be exploited on wider range of applications since it can process large number of samples.

KIN Algorithm:

Input: D: Dataset = $\{D_t, D_{t+1}.D_{t+2},\ldots,D_{t+j}\}$

$\qquad D_t = x_j \in X; y_j \in Y$

$\qquad X = \{x_0, x1,\ldots.x_j\}$

$\qquad Y = \{y_0, y1,\ldots,y_j\}$

Assume each D requires processing time:

$T = \{T_t, T_{t+1}, T_{t+2},\ldots.T_{t+j}\}$

For t = 1,2,…. Do

1. Obtain D_t for each learner C

Where C : Classifier = $\{C_N,C_I,C_K\}$

2. Apply C on D_t to produce individual hypothesis.

i.e.$H_K = C_K (D_t)$

$\qquad H_I = C_I (D_t)$

$H_N = C_N (D_t)$

3. Calculate voting weights of each H using majority voting

$H^* = MajV (, H_K,H_I,H_N\}$

Output: Returns H^* : Final Hypothesis

IV. EXPERIMENTS AND RESULTS

In the first experiment, performance of proposed ensemble system (KIN) has been compared with each base incremental learners (KStar, IBK,NNGe) for credit card risk assessment. Table 1 illustrates

this comparative result with respect to accuracy of learner, time required to build/train the learner, and time required to evaluate the model. By observing values recorded in Table 1, it can be pointed out that KIN performs best among all other (individual) incremental base learners. To compute the correctly classified instances by the system, the entire training dataset was partitioned into ten equal-sized and mutually-exclusive chunks and for each chunk the classifier was trained on the integration of another chunks.

Table 1.Comparison of the KIN technique with existing incremental algorithms

	KIN	KStar	IBK	NNGe
Accuracy (In percentage)	74.1	69.5	72	70.5
Time To Build Classifier (In seconds)	1.9	0.01	0.01	3.36
Time To Evaluate Classifier (In seconds)	9.5	10.2	1.5	5.1

Table 2. Comparative performance of KIN with existing incremental algorithms (detail accuracy by Class: Good)

	KIN	KStar	IBK	NNGe
TP Rate	0.879	0.813	0.81	0.841
FP Rate	0.58	0.583	0.49	0.613
Precision	0.779	0.765	0.794	0.762
Recall	0.879	0.813	0.81	0.841
F-Measure	0.826	0.788	0.802	0.8

Table 3. Comparative performance of KIN with existing incremental algorithms (detail accuracy by Class: Bad)

	KIN	KStar	IBK	NNGe
TP Rate	0.42	0.417	0.51	0.387
FP Rate	0.121	0.187	0.19	0.159
Precision	0.597	0.488	0.535	0.511
Recall	0.42	0.417	0.51	0.387
F-Measure	0.493	0.45	0.522	0.49

Then, cross validation was run on each of them. This paper also shows comparative performance of KIN with detail accuracy by class for credit risk evaluation with respect to both classes, i.e. good and bad in table 2 and table 3 respectively. This illustrates extensive analysis of the credit risk assessment system using KIN technique.

While determining the performance prediction of any algorithm the nature and size of the data plays an important role. German credit dataset is the standard dataset downloaded from UCI repository consisting of 1000 samples, 2 classes and 24 attributes and all instance values are numerical. Statistical nature of this data has been tested and it is found that data is discrete. Data is useful in describing the behavior of an algorithm. In recent years, the concept of integrating learners is proposed as a new approach for the achievement of better accuracy and thus great performance. Moreover, most of the ensemble based system work in batch mode. But here, the proposed ensemble (KIN) operates incrementally and the algorithms in the systems are also picked such that they could operate incrementally and maintain proper diversity in the ensemble. Hence the system is more inclined towards dynamic nature and not the static one. In the next experiment, comparison of

existing state- of- art classifiers for credit risk evaluation is performed with proposed ensemble (KIN).

It has been recorded that KIN system got best accuracy among them, Table 4 shows the detail analysis of this experiment. Additionally, the graph shown below in figure illustrates comparative analysis of KIN with well-known classifiers.

Table 4.Comparison of KIN with popular classifiers for credit risk assessment

	KIN	J48	SVM	LWL	AdaBoost	Bagging
Accuracy (In percentage)	74.1	70.5	70	70	69.5	73
Time To Build Classifier (In seconds)	3.1	0.24	3.12	0.1	0.31	0.78
Time To Evaluate Classifier (In seconds)	9.5	0.9	1.6	1.3	1.0	1.3

CONCLUSION

Credit card threat assessment is an important issue in financial systems and the users associated with such systems since the frequency of credit card frauds in increasing. Several machine learning techniques are applied for assessment of credit risks. Incremental and online nature of credit data lead the way towards dynamic systems and created new challenges for knowledge acquisition. Ensemble based system is an intelligent approach for improvement in the performance of data mining algorithms. The proposed ensemble based framework (NIK) in this paper performs well forevaluation of credit risks and this is described with several experiments.

Using incremental algorithms and proposed ensemble in several synthetic and artificial datasets and statistical testing is critical area of future work and reasonable training and evaluation time in most of the cases.

REFERENCES

[1] Wang, Gang, and Jian Ma, 2012 "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machines" Elsevier- Expert Systems with Applications

[2] An incremental ensmeble of classifiers, S.B. Kotsiantis, Published online: 11 March 2011, Springer Science and Business Media B. V. 2011

[3] Woźniak, Michał, Manuel Graña, and Emilio Corchado., 2014, "A survey of multiple classifier systems as hybrid systems."

John G. Cleary, Leonard E. Trigg, 1995, K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114

[4] Cover, Thomas, and Peter Hart.,1967, "Nearest neighbor pattern classification." Information Theory, IEEE Transactions on 13, no. 1

[5] Roy, Sylvain,2002 "Nearest neighbor with generalization." Christchurch, New Zealand .

[6] D. Aha, D. Kibler ,1991. Instance-based learning algorithms. Springer, Machine Learning. M:January 1991, Volume 6, Issue 1, pp 37-66

[7] N. Littlestone and M.Warmuth,1994 "Weighted majority algorithm," Inform. Comput., vol. 108, pp. 212–261

[8] PallaviKulkarni and Roshani Ade., 2014 Article: Prediction of Student's Performance based on Incremental Learning. International Journal of Computer Applications 99(14):10-16

[9] Ade, Ms RR, et al.,2013, "Methods For Incremental Learning: A Survey "International Journal of Data Mining & Knowledge Management Process 3.4 (2013)

[10] PallaviKulkarni and Roshani Ade, 2014," Incremental Learning From Unbalanced Data With Concept Class, Concept Drift And Missing Features: A Review", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.6