Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 12, December 2013, pg.229 – 238

SURVEY ARTICLE

SURVEY OF CLASSIFICATION RULE MINING TECHNIQUES FOR IDENTIFYING DISEASE CAUSE AND DIAGNOSIS

K.S.Thirunavukkarasu [1], Dr. S.Sugumaran [2]

[1] Ph.D Research Scholar, Manonmaniam Sundaranar University, Assistant Professor, Department of computer science, Nehru Memorial College, Trichy, TamilNadu, India
[2]Associate Professor, Department of computer science, Erode Arts and Science College, Erode, TamilNadu, India
[1] thirukst@gmail.com; [2] prof_sukumar@yahoo.co.in

Abstract-Classification is a supervised learning technique. Classification arises frequently from bioinformatics such as disease classifications using high throughput data like microarrays. Classification rule mining classifies data in constructing a model based on the training set and the values or class labels in a classifying attribute and uses it in classifying new data. Currently, a various modeling techniques are detailed for data mining. The details of data mining and machine knowledge in related and network domains are dependent and comparatively distributed. The technique particularly achieves the statistical belief among occurrences in order to enhance classification accuracy. An attention on dependencies is made where the ability to draw classification accuracy is affected in improving performance of the model.

Data partitioning approaches such as bagging and boosting are greatly handled in multiple classifier systems to improve classification accuracy. Most current data stream classification techniques fails in one essential aspect of stream data i.e. arrival of a class. So, a data stream classification method that merges a class detection system into traditional classifiers is enabled. The automatic detection of classes before true labels arrive is detected. The problem of data stream classification, where the data appear in an abstractly limitless stream and the chance to analyze each record is briefed. The searching of a training set accurately and classifying is difficult while considering a large data set. Even with the classified data set the accuracy of the classification is inefficient with error rates. This paper presents classification based on shared information for diagnosing disease. The medical dataset is analyzed with stroke disease reducing error rates providing classification accuracy. This paper also reviews certain data mining papers on classification rule for disease diagnosis patterns.

Keywords— Bayesian classifier; Rule mining; Random forest; Classification; Review

I. INTRODUCTION

In data mining the data appear in abstractly limitless stream for classification of data stream. The problem of data stream classification, where the data enter in an unreal infinite stream and the chance to evaluate each record is briefed. The

problem is solved with the existence of stream classification algorithm. The stream classification algorithm runs in payback O (1) time which is capable of managing irregular arrival of labeled records. The algorithm is also able to adapt its parameters to respond changing in class boundaries of data stream. In addition, the algorithm is capable of judging the quality of models updated on deploying unlabeled records. Based on the unlabeled records the additional requirement of labeled records is decided.

A major concern when incorporating large sets of multiple n-gram features for classification is the presence of noisy, irrelevant, and redundant attributes. These responsibilities often make it crucial to control the augmented discriminatory potential of extended feature sets. To handle a large set of semantic information a classification rule is used which depends on multivariate text characteristics selection method called Feature Relation Network (FRN). Also leverages the acceptable relationships between n-gram features. FRN is predetermined to thoroughly enable the addition of extended sets of heterogeneous n-gram features for enhanced sentiment classification. FRN is more efficient compared to univariate, multivariate, and hybrid feature selection methods. FRN increases classification accuracy by choosing attributes regardless of the feature subset sizes. In addition, FRN supports more feature selection process by combining semantic information providing computationally efficient method compared to multivariate and hybrid techniques. Data partitioning approaches such as bagging and boosting are widely handled in multiple classifier systems. The approach to face multiple classifiers shows a higher potential in enhancing classification accuracy. The study is related to training data distribution analysis and its effect on the behavior of multiple classifier systems. Various feature-based and class-based measures are used to determine analytical features of the training partitions. The various types of training partitions along with different distribution are produced and evaluations on huge training partitions are made. Then, the training partitions and their effect on the achievement of the system are determined by utilizing the feature-based and class-based measures. The partitioning method termed as Clustering, Declustering, and Selection (CDS) is developed based on the measure analysis.

The existences of class are unnoticed in most current data stream classification techniques which is an important feature of stream data. So, a novel class detection mechanism is developed that links class detection into traditional classifier. The integration of class detection mechanism enables automatic detection of classes before the arrival of class instances. The difficulty in detection of class is complex with the presence of concept drift as hidden data distribution affect the progress of stream. Estimation on class is made to find the number of instances belonging to class. The estimation involves a more test instances in the classification model for discovering similarities among the instances. A maximum acceptable wait time is set as a time constraint to classify a test instance. In addition, most current stream classification methods are accessed immediately after the classification of data point. Practically, a time delay is seen in acquiring the true label of a data point as manual labeling is time consuming. A fast and correct classification decisions are made on real benchmark data.

A time consuming temporal data mining method is required for better performance in classification mining. So, an approach for temporal data mining based on classification rule for easy understanding of human domain experts is generated. Generally, time series are disintegrated into short segments, and short-term trends of the time series within the segments. The disintegrated time series short segments like average, slope and curvature are segmented by polynomial models. The classifiers determine short sequences of flow in consecutive segments with their rule premises. The resultants slowly allot an input to a class. The time series are assigned to originate with the productive time series model as classifier detects the anomalies.

The time series pass provides a faster modeling in segmenting and piece wising polynomial models. The classification rule approach is appropriate to problems with rough timing constraints. The theoretical foundations for classifier, containing a distance measures for time series demonstrate the efficient functioning of classification rule. Classification mining classifies data in diagnosing disease based on the medical data set. The class labels in a classifying attribute provides the cause for disease and existence in classifying diagnoses stroke disease. Classification rule mining techniques for identifying disease cause and diagnosis work aims in:

- Diagnosing disease through simple classification tree methods on improving search for dividing arithmetic attribute points.
- Considering classification algorithm in enhancing classification accuracy for analyzing stroke disease in terms of error rates and the connected attribute size.
- Developing a feature selection based approaches with classification rule to reduce space dimension of classifiers on improving classification.

This paper is organized as follows: Section II discusses classification rule for disease cause and diagnosis, Section III shows the study and analysis of the existing classification rule techniques in data mining, Section IV identifies the possible

comparison between them and Section V concludes the paper, key areas of research is given as making use of classification rule mining in disease cause as well as diagnosing disease and designing new algorithms and systems for efficient classification.

II. LITERATURE SURVEY

The problem of data stream classification is the arrival of data in an abstractly immeasurable stream and the chance to determine each record is briefed [1]. An online stream classification algorithm executing in amortized O (1) time controls the irregular appearance of labeled records. The algorithm judge internally on the basis of quality of updates models. The models are updated from the unlabelled records to identify the inclusion of labeled records. The stream-classification algorithm handles multiple target classes. A new technique for temporal data mining based on classification rules with human domain experts easy understating is enabled [5]. Generally, time series are disintegrated into short segments, and short-term trends of the time series within the segments. The disintegrated time series short segments like average, slope and curvature are segmented by polynomial models. The classifiers determine short sequences of flow in consecutive segments with their rule premises. The resultants slowly allot an input to a class. The time series are assigned to originate with the productive time series model as clas]sifier detects the anomalies. The time series pass provides a faster modeling in segmenting and piece wising polynomial models. The method is appropriate to problems with harsh timing constraints. A large data set source classification is a limitation. A test chain classification is enabled to overcome the limitation [17]. The chain classification for large dataset classifies data based on the knowledge in the extending areas of adjacent scenes. The basic idea was to classify one dataset initially where the best truth data is present. Then to classify the adjacent dataset using classification of the initial overlap set as a training data. The arrival of novel class aspects are handled in the data stream classification method [2]. The data stream classification technique binds a novel class detection mechanism into traditional classifiers. The integration implements automatic detection of novel classes before the occurrence of true labels novel class instances. More time consumed for recognizing similarity instances in a class model. In addition, most existing stream classification techniques access the true label of a data point speedily after the data point is classified.
‘

The ranking and classification are combined to provide more accurate determination of a heterogeneous information network [16]. Highly ranked objects within a class play a vital role in classification. Similarly a class membership detail is essential for evaluating a quality ranking over a dataset. The efficiency of collective classification models is improved depending upon the amount of class labels information present [15]. An availability change is noticed with the increase in test set labels for the relative performance of statistical relational models. The Data partitioning methods like bagging and boosting are greatly utilized in multiple classifier systems. The classification accuracy is highly improved based on data partitioning methods. The analysis of training data distribution and its effects on the activities of multiple classifier systems is studied [3]. Various feature-based and class-based measures are used to determine the statistical functioning of the training partitions. The measures assess the importance of different types of training partitions.

The modeling techniques for data mining and network domains are independent and identically distributed. The techniques particularly attain the statistical dependencies among instances in order to enhance classification accuracy. The ability to draw accurate result for the performance of the models depends on the same dependencies. The complicated connection patterns and attribute dependencies in relational data disrupt the assumptions of many conventional statistical tests. The work of network classification within network and the algorithm for models concludes significantly different levels of performance [12]. The data mining methods on derivatized tandem datasets shows high classification accuracies [14]. Non-derivatized screening based methods are negotiated in machine learning approaches. Three data mining methods, namely C4.5 decision trees, ridge logistic regression and logistic regression are established. Threshold optimization method evaluates the applicability in terms of diagnostic support tool.

The existence of noisy, irrelevant, and redundant attributes in integrating large set of diverse n-gram features for sentiment classification is avoided [4]. A rule-based multivariate text feature selection method called Feature Relation Network (FRN) control the existence of noise and irregularity attributes. FRN deal with semantic knowledge and also influences the syntactic relationships between n-gram features. FRN is designed to sufficiently allow the addition of continued sets of heterogeneous n-gram features for improved sentiment classification. An online interpretation of the popular learning with local and global consistency method (OLLGC) is presented in [20]. The structural properties of a graph classification are proved.

The query classification is focused. Query classification is important for understanding the user determined both in Web search and in online publicity. A robust method for classifying non-English queries into an English taxonomy, using an existing English text classier is designed along with the-shelf machine translation systems [13]. The Bayesian classifier is a traditional classification technique. The programming Bayesian classifier into SQL is focused [7]. Two classifiers: Naive Bayes and a classifier depending on class decomposition using K-means clustering are established. Two complementary tasks: model

computation and scoring a data set are examined. The work transforms equations into efficient SQL queries and establishes several query optimizations.

A study on machine learning methods to email prioritization into discrete levels and comparison is made between the ordinal regressions and a classifier cascade is presented [8]. But a cascade of SVM classifiers has considerably performed well the ordinal regression for email prioritization. SVM regression performance is well better when compared to classifiers. An embedded approach is presented that which concurrently chooses relevant features at the time of classifier construction by disciplining each feature's use in the dual formulation of support vector machines (SVM). The kernel-penalized SVM (KP-SVM) approach optimizes the shape of anisotropic RBF Kernel removing features which has no importance for classifier [11]. Additionally, KP-SVM utilizes an explicit stopping condition that which avoids the removal of features that would harmfully affect the classifier's performance. The lazy associative classifiers (LAC) are improved by the usage of an MDL-based entropy minimization method [9]. Well calibrated classifiers are those which provide accurate evaluations of class membership probabilities. Investigation is made on crucial applications where such characteristics like accuracy and calibration are relevant and demonstration is also made depending on that the LAC performs well when compared to other classifiers, such as SVMs, Naive Bayes, and Decision Trees even though these classifiers are calibrated. LAC also has the ability to incorporate reliable predictions in order to improve training, and the ability to renounce from doubtful predictions.

A framework for discriminative sequence classification is linear classifiers that work precisely in the explicit high-dimensional predictor space of every subsequence in the training set. A gradient-bounded coordinate-descent algorithm is employed for selecting discriminative subsequences but without enlarging the whole space [8]. The framework is applicable to a wide range of loss functions that includes binomial log-likelihood loss of logistic regression and squared hinge loss of support vector machines. The task of user activity classification in microblogs is denoted by the author, where users can distribute short messages and control social networks online [10]. The significance of modeling a user's individuality is identified and that of developing opinions of the user's friends for correct activity classification. A new collaborative boosting framework includes a text-to-activity classifier for each users. It also involves a method for collaboration among classifiers of users having social connections.

The Prototype-based Domain Description rule (PDD) one-class classifier is introduced [19]. PDD is an adjacent neighbor based classifier and it accepts objects depending on their adjacent neighbor distances in a reference set of objects known as prototypes. For choosing suitable prototype set, the PDD classifier must be equivalent another adjacent neighbor depending on one-class classifier called as the NNDD classifier. A guide on the future research is provided in the analysis and facilitate knowledge accumulation [6] is facilitated. Creation regarding the application of data mining methods in Customer Relationship Management is also provided.

III. TECHNIQUES OF CLASSIFICATION MODEL

Classifiers in current classification algorithms need highly huge amounts of data for training. The algorithms should utilize less data for training as possible to reduce the classifier time. Classification rule depicts the information in the form of IF-THEN rules. A rule is generated based on the criteria for each path from the root to a leaf. Each attribute value combination with a path models a conjunction. The leaf node employs the class classification. Classification rule to improve the searching and classification accuracy.

. The classification algorithm is applied to the medical data set. A classifier identifies the stroke disease with the conditions and differentiates it from all. The classification algorithm is designed based on the classification tree. The classification tree method is utilized for diagnosing disease and also improving search for dividing arithmetic attribute points. The Classification tree construction is build from top-down recursive methods like divide and conquer patterns. Initially start the root attributes with the medical data set example categories. The samples are partitioned repeatedly depending upon the selected attribute values. Test attributes are chosen on the criteria of feature based classification measures. The dividing points are searched for a given node belongs to the same class. Final results should give an empty set for further classifying. Major votes are included for classifying the leaf leaving no samples at the end.

A. Classification Using Streaming Random Forests.

The essential idea is to amalgamate the techniques used for building decision trees from streams, and for building Random Forests to create an efficient and accurate algorithm, that can be updated in an incremental way since the classification model is an ensemble.

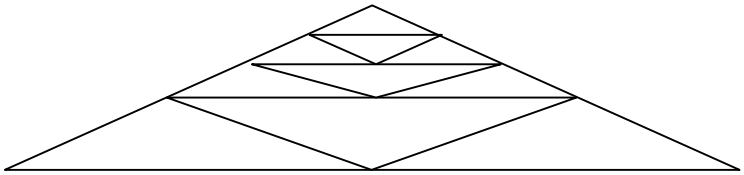

Fig 2 Classification in Tree Form

Fig 2 describes the instructions on how to extend the continuous lines of the bigger triangle so as to make a subsection on trees for further classification. The initial phase introduces the way in which the streaming decision tree construction is merged with the Random Forests algorithm to prepare an incremental stream classification algorithm that handles with classification accuracy comparable to standard classification algorithms. The second phase shows how the algorithm extended to handle concept drift in the stream that is new labeled records that imply a different set of decision boundaries among the classes. The algorithm of final phase is a self-adjusting algorithm that handles using an entropy-based change-detection technique. The key feature of this phase is that the algorithm is now intelligent to choose whether the current model is ready for deployment or not, when the number of labeled records is not enough to completely build/update the model.

// Algorithm for building and updating
1: While more trees to build
2: Call Build tree
3: If block of label records ends
4: Number of trees calls then Evaluate Forest Func
5: Initial Building uses the Current Forest for Classification
6: Reset Evaluation Set
7: End If
8: Continue building trees when a new block of labeled records arrive
9: If Concept drift is detected
10: Reset Evaluation set
11: End If
12: End While

As in the original building phase, when the algorithm completes an evaluation phase and the forest is ready for consumption, the classification error for the evaluation set is designed, the evaluation set is unoccupied, and the algorithm enters a test phase to simulate the deployed classification error. The evaluation set is also vacant when a concept modify is detected. This ensures that the algorithm evaluates the forest on data with the equivalent data distribution as the data on which the forest is deployed. If the forest is not prepared for operation, the algorithm waits for the arrival of the next block of labeled records, and resumes the update process. The above classification tree algorithm is used to enhance the search with the training data set. Initially the classification tree is build using the while loop. The classification rule for the given training data set is evaluated based on block of labeled medical data records if condition. If the existence of disease cause is recognized then diagnose the stroke disease. If negative sign for disease is seen then it indicates patient is safe. Restore old tree. Close the 'if' loop. Reset the evaluation set if new medical data set is arrived. Close the while loop. End the process. As with any tree ensemble classifier, it produces a number of binary decision trees and classifies the class of each new record using the group of class predictions from the set of trees [1]. An incremental stream classification algorithm attains high classification accuracy, even for multiclass classification problems. The algorithm combines the ideas of streaming decision trees and Random Forests. The algorithm is intricate, so introduce its explanation in three phases. The most classification fails in multiple target class managing. But stream classification algorithm solves multiple targeted classes. The cause for the disease is recognized using classification tree to improve efficiency in the search with multiple targeted classes. With the detection of multiple targeted classes an improvement in classification accuracy is required.

B. Classification and Class Detection in Concept-Drifting Data Streams.

Data stream classification techniques ignore one imperative aspect of stream data such that the arrival of a novel class. In order to decide whether an example belong to a narrative class, the classification model sometimes needs to wait for more

test instances to discover similarities among those instances. A greatest allowable wait time T_c is imposed as a time constraint to classify a test instance.

Input: L: Current collection of best M classifiers, x_j: test example, buf: buffer holding provisionally deferred instances
Output: Immediate or deferred class prediction of x_j
1: fout true
2: if Foutlier(L; x_j) = false then
3: y_0= majority-voting(L; x_j) //classify immediately
4: fout false
5: End if

The above steps are used for classification. The test instance x_j is a Foutlier. So, if x_j is not a Foutlier, we classify it instantly using the collection voting. Recall that a novel class instance must be a Foutlier. Conversely, a Foutlier is not necessarily an existing class instance. Therefore, carry out additional analysis on the Foutliers to decide whether they really belong to novel class. The below Fig 3 describes the classification process when the rule is generated.

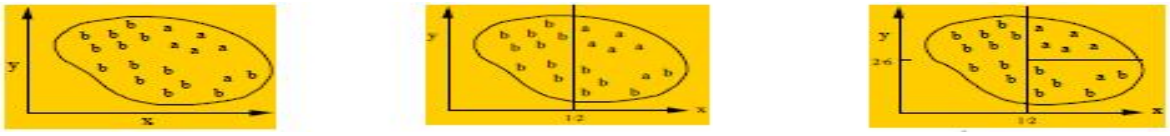

Fig 3 Classification when generating the rule

In addition, most existing stream classification approaches suppose that the true label of a data point can be accessed instantly after the data position is classified. In reality, a time delay T_l is involved in attain the true label of a data point because physical labeling is time consuming but the classification technique is not applied to the network traffic. Besides, the data stream classification problem under dynamic feature sets is not addressed.

C. Data Partitioning for Training Multiple Classifier Systems

In a deviation from the traditional approach, the use of Multiple Classifier System (MCS) works to improve classification accuracy [3]. Classifier use a wrapper model for attribute selection, where the accuracy of a target classifier is used as an evaluation metric for the classification of a particular feature subset. MCS depends on many different factors. The original data to training, X_i train, and test sets is divided. The goal was to produce partitions with dissimilar degrees of class diversity. As a result, training set X_i train was again divided in such a way that the first set of training partitions S_i contained only one class. If the number of classes was smaller than the number of partitions, larger classes were distributed among two or more subsets.

Furthermore, if the number of classes was larger than the number of partitions, two or more classes were grouped into one training subset, depending on the size of the classes. These sets remained the baseline for the generation of succeeding set of solutions. By randomly selecting a certain amount of instances X_{rand} from the training data X_i train and distributing these among initial subsets, new subsets S_i were generated. The size of training partitions remained impartial all the way through the experiments. Additional sets of training partitions were generated by increasing the size of X_{rand} and distributing. It is important to note that, M, b, and r are user-defined values. M and b control the number of training divider generated all through the experiments and r specifies the size of random instances distributed in each training partition set.

//Data Generation Scheme
Inputs: X original data, M and b: are predefined integer values, r: is a predefined integer between 1 and 10, m: number of partitions
Output: An initial set of training partitions generated in iteration by manipulating.
1: for i = 1:b, generate X_i
2: Train from X, generate S_i
3: t=2
4: While frac<=100
5: For n = 1 : M
6: Generate X_{rand} by randomly selecting frac%
7: t=t+1
8: End for
9: frac = frac + r

10: End while
11: End for

The original and the nearly all famous observation was that the MCS error was correlated with feature-based and class-based measures, suggesting that there exists a set of training partitions for which the MCS accuracy is at its best. Another watching was that the construction of numerous classifiers on partitions encloses diverse classes resulted in higher classification accuracy. Finally, a larger distance among training partitions increased the probability of error in the ensembles. It does not develop a sophisticated methods and measures for training data subsets with overlaps. It fails in combining the filter-based data partitioning approach with a wrapper-based method.

D. Selecting Attributes for Sentiment Classification.

Recursive feature based measures uses a wrapper model based on an SVM classifier [4]. Multi-classifier class combines the classification mining approach to feature based measures. In case of intrusion traditional detector undetected interruption are faced. Associative classifier enriches disease cause detection. The associative classifier is collection of high quality classification rules, which are developed from highly assured events that follow close dependencies among events.

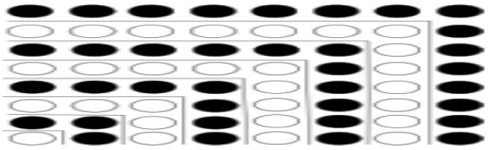

Fig 4 Diagrammatic Classification

Fig 4 contains a message on the subsequently probable step to carry out, but in the first case the instruction given is on how to add the next row of discrete white spots. A rich set of n-gram features spanning many fixed and variable n-gram categories coupled the extended feature set with a feature selection method competent of professionally identifying an enhanced subset of n-grams for opinion classification. The Feature Relation Network (FRN) is a rule-based multivariate n-gram feature selection technique that professionally eliminates redundant or less useful n-grams, allowing for more effective n-gram feature sets.

FRN also incorporates semantic information consequent from existing lexical resources, enabling augmented weighting/ranking of n-gram features. FRN results attained are appropriate for other text classification problems, where semantic information is obtainable (e.g., topic, affect, and style classification). We also intend to explore additional potential feature relations. The additional feature occurrence measurements are not added. Exchange semantic weighting mechanisms and development of hybrid feature selection methods that incorporate FRN in conjunction with other multivariate selection techniques are not explored.

IV. COMPARISON OF CLUSTERING TECHNIQUE & SUGGESTIONS

In order to compare the running time of the different classification algorithm, set of record classes are taken to perform the experiment. The initial metric is the running time, is defined as the time taken to perform the classification process on multi dimensional data. The second performance metric error rate is the number of bit errors occurred on data stream while diagnosing the disease. The comparison takes place on existing Classification system Using Streaming Random Forests (SRF), Classification in Concept Drifting Data Streams under Time Constraints (CDDS), Filter-Based Data Partitioning for Training Multiple Classifier Systems (MCS) and Selecting Attributes for Sentiment Classification Using Feature Relation Networks (FRN).

No. of classes		**Running Time (ms)**		
	Classification in CDDS	**FRN**	**MCS**	**SRF**
10	991	857	802	756
20	981	851	794	758
30	983	874	817	774
40	995	893	816	783
50	984	909	824	782
60	983	895	826	791
70	974	876	838	796

Table 4.1 Tabulation for running time on different classification technique

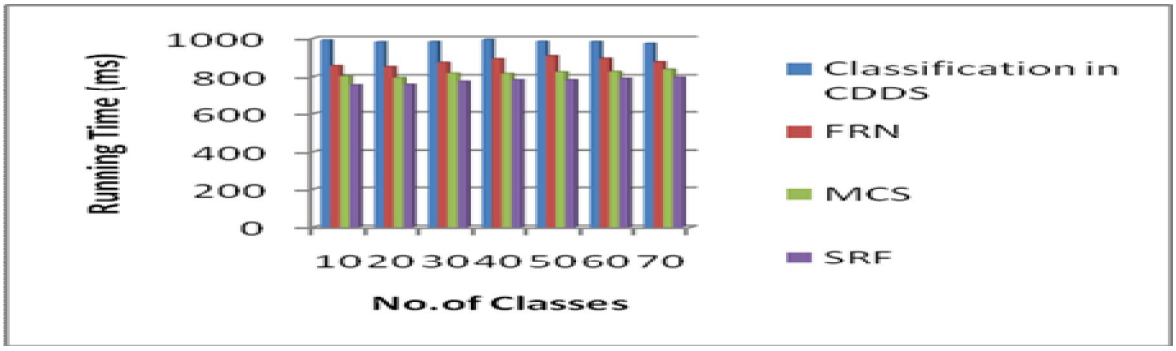

Fig. 4.1 Running time on different classification Technique

Fig 4.1 describes the running time based on the record classes. The running time is measured in terms of milliseconds (ms). As the class's increases, running time is reduced in the classification in CDDS. The experiment shows that classification in CDDS greatly brings down the time while performing the execution when compared with the FRN, MCS, and SRF. It can be seen that the Classification in CDDS shows conspicuous advantage.

Classification in CDDS is 10 – 15 % lesser delay time taken when compared with the FRN and 15 -20 % lesser running time taken when compared with the MCS. Classification in CDDS is approximately 25% lesser running time taken for when compared with the SRF.

Classification Technique	**Classification Error rate (%)**
MCS	0.1649
Classification in CDDS	0.2587
SRF	0.3578
FRN	0.4518

Table 4.2 Tabulation for Error rate of different technique

The above table (Table 4.2) describes the error rate of the MCS, Classification in CDDS, SRF and FRN. Error percentage of MCS [3] is lesser when compared to the all other existing classification technique. The raw data illustrating the effects of error rate on different techniques are shown in Fig. 4.2. Error rate of MCS is decreased using the partitions with dissimilar degrees of class diversity. The variance is 10 – 15 % lesser in MCS when compared with MST based Clustering method.

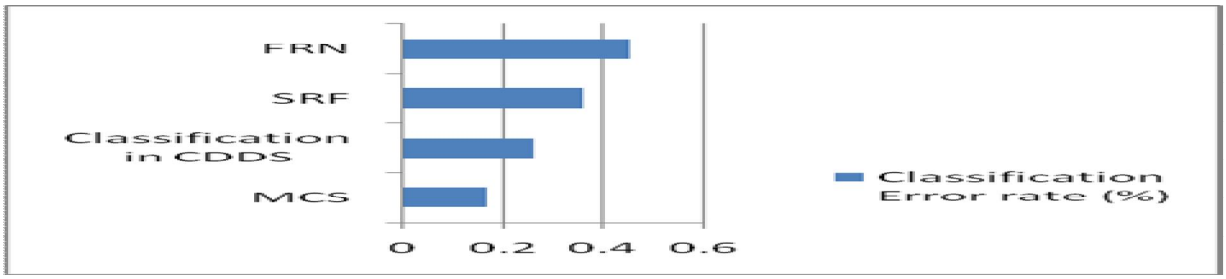

Fig 4.2 Classification Error rate of different technique

Fig 4.2 demonstrates the error rate of different techniques. The usage of clearance in the MCS decreases the classification error rate when compared with the other existing classification algorithms. Classification rule improves the searching and classification accuracy reducing error rates. The cause for the disease is identified with the class labels in a

classifying attributes. Existing papers has reviewed the potential of the classification algorithms. More classification rule mining techniques developed feature-based and class-based measures for searching but the classification accuracy is not improved in diagnosing the disease cause. A lazy associative classifier does not follow the classification accuracy maximization paradigm i.e., it commonly portray the training data. Survey has reviewed the searching efficiently using classification algorithm for diagnosing the disease cause with improved classification accuracy. A survey and contribution on classification rule mining techniques are available from recent research.

The proposal classification rule mining technique for diagnosing the disease cause as a way forward to

- Searching efficient with dividing arithmetic attributes points for diagnosing disease using classification tree methods.
- Offers classification algorithm with improved classification accuracy reducing error rates for analyzing stroke disease.
- Reduced Space dimension classifier due to development of feature selection based approaches with classification rule.

Classification rule improves the searching and classification accuracy reducing error rates. The cause for the disease is identified with the class labels in a classifying attributes. The presence of the stroke disease existence is diagnosed with the classification rule algorithm. Based on the medical training data set for diagnosing disease is classified by classification mining. Classification accuracy is enhanced by reducing error rates for diagnosing stroke disease. Classification rule including feature selection based technique with less space classifier improves classification mining.

V. CONCLUSION

Discussion about existing Classification system Using Streaming Random Forests, Classification in Concept Drifting Data Streams under Time Constraints, MCS and Selecting Attributes for Sentiment Classification Using Feature Relation Networks. Streaming random forests algorithm productively switch concept changes using entropy based conception drift detection technique. It quickly records the new expected classification accuracy after the changes are presented in the stream. It also dynamically adjusts its parameter based on the data seen.

Existing data stream classification techniques assume that totality number of classes in the stream is fixed. Therefore, instances belonging to a novel class are misclassified by the existing techniques. The experiment show the result of how detect novel classes automatically even when the classification model is not trained with the novel class instances. Novel class detection becomes more demanding in the attendance of concept-drift. Existing classification rule mining algorithm for enhancing searching feature based measures is ignored in implementing diagnosis disease cause. This is because the focus is made only on searching and not on feature based diagnostic. In addition the accuracy is not considered so classification accuracy handling is focused. A review shows that classification rule mining is the growing need of data mining and with high classification accuracy. Hence the classification rule mining technique helps in satisfying the efficient diagnosis of stroke disease cause.

Surveillance was that the building of multiple classifiers on partitions containing diverse classes resulted in higher classification accuracy. Finally, a larger distance among training partitions increased the probability of error in the collection. However, these tests have given us a strong indication of the number of underlying classification in the data, as well as how each method performs under various conditions. The extensive experiments evaluate the relative performance of the various classification algorithms and combinations. The result shows that the classification technique outperforms consistently over a wide range of experimental parameters.

REFERENCES

[1] Hanady Abdulsalam, David B. Skillicorn and Patrick Martin, "Classification Using Streaming Random Forests", IEEE Transactions On Knowledge And Data Engineering, January 2011.

[2] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", IEEE Transactions On Knowledge And Data Engineering, June 2011.

[3] Rozita A. Dara, Masoud Makrehchi and Mohamed S. Kamel, "Filter-Based Data Partitioning for Training Multiple Classifier Systems", IEEE Transactions On Knowledge And Data Engineering, April 2010.

[4] Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions On Knowledge and Data Engineering, March 2011.

[5] Dominik Fisch, Thiemo Gruber, and Bernhard Sick, "Swift Rule: Mining Comprehensible Classification Rules for Time Series Analysis", IEEE Transactions On Knowledge And Data Engineering, May 2011.

[6] E.W.T. Ngai, Li Xiu and D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", ELSEVIER Science Direct 2009.

[7] Carlos Ordonez and Sasi K. Pitchaimalai, "Bayesian Classifiers Programmed in SQL", IEEE Transactions on Knowledge and Data Engineering, January 2010.

[8] Georgiana Ifrim and Carsten Wiuf, "Bounded Coordinate-Descent for Biological Sequence Classification in High Dimensional Predictor Space", ACM 2011.

[9] Adriano Veloso, Wagner Meira Jr, Marcos Gonçalves, Humberto M. Almeida, Mohammed Zaki, "Calibrated lazy associative classification", ELSEVIER Science Direct 2011.

[10] Yangqiu Songx, Zhengdong Luz, Cane Wing-ki Leungz and Qiang Yang, "Collaborative Boosting for Activity Classification in Microblogs", ACM 2013.

[11] Sebastián Maldonado, Richard Weber and Jayanta Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines", Science Direct, ELSEVIER 2011.

[12] Jennifer Neville, Brian Gallagher, Tina Eliassi-Rad and Tao Wang, "Correcting evaluation bias of relational classifiers with network cross validation", Springer, Knowledge Information System, 2012.

[13] Xuerui Wang, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski and Bo Pang, "Cross-Language Query Classification using Web Search for Exogenous Knowledge", ACM 2009.

[14] Tim Van den Bulcke, Paul Vanden Broucke, Viviane Van Hoof, Kristien Wouters, Seppe Vanden Broucke, Geert Smits, Elke Smits, Sam Proesmans, Toon Van Genechten and François Eyskens, "Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data", ELSEVIER 2011.

[15] Rongjing Xiang and Jennifer Neville, "Understanding Propagation Error and Its Effect on Collective Classification", IEEE International Conference on Data Mining, 2011.

[16] Ming Ji, Jiawei Han and Marina Danilevsky, "Ranking-Based Classification of Heterogeneous Information Networks", ACM 2011.

[17] Jan Knorn, Andreas Rabe, Volker C. Radeloff, Tobias Kuemmerle, Jacek Kozak and Patrick Hostert, "Land cover mapping of large areas using chain classification of neighboring Landsat satellite images", ELSEVIER Science Direct 2009.

[18] Shinjae Yoo, Yiming Yang and Jaime Carbonell, "Modeling Personalized Email Prioritization: Classification-based and Regression-based Approaches", ACM 2011.

[19] Fabrizio Angiulli, "Prototype-based Domain Description for One-Class Classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.

[20] Quanquan Gu, Charu Aggarwal, Jialu Liu and Jiawei Han, "Selective Sampling on Graphs for Classification", ACM 2013.

