Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7642710

DATA MINING USING SUPERVISED INSTANCE SELECTION (SIS) FOR BETTER CLASSIFICATION ACCURACY IN ARTIFICIAL NEURAL NETWORKS

S. Srinivas Reddy and Dr. Rajeev G. Vishwkarma
Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University, Indore
(M.P.) - 452010, India
Corresponding Author Email : srinivasreddy.sarvigari@gmail.com

ABSTRACT: The semi-supervised learning techniques use abundant unlabeled data for helping to learn a better classifier if the number of instances is very less. A basic technique is to choose and label the unlabeled instances that the present classifier has higher classification confidence for enlarging the labeled training set and the updating the classifier, which is mostly utilized in two different paradigms of semi-supervised learning namely: co-training and self-training. But the actual labeled instances will be more reliable compared to self-labeled instances which would be labeled by a classifier. If unlabeled instances are assigned to wrong labels then the classification accuracy of classifier might be jeopardized. In this paper, a new instance selection technique is presented based on real labeled data. This will consider present classifier performance on unlabeled data as well as its performance only on real labeled data. In every iteration, this utilizes the accuracy changes in newly learned classifier over original labeled data as a criteria for deciding either the chosen most confident unlabeled instances would be accepted by further iteration or not. The experiments will be conducted in co-training as well as self-training while using Naïve Bayes (NB) as a base classifier. The results show that, SIS will significantly improve accuracy and classification of self-training and co-training. From results it can observe that it will improve the accuracy, Precision, Recall and F1 score compared with semi-supervised classification method.

KEYWORDS: Supervised Instance Selection (SIS), Data mining, Meta-learning, Algorithm selection.

I.INTRODUCTION

Data reduction methods have been regularly utilized for lessening preparing time and capacity necessities of the classifiers. Data reduction is comprised of data 3D square accumulation, data pressure, include/characteristic selection and instance selection. Data 3D shape and highlight selection helps in decreasing the dimensionalities of data to be mined and hence diminish the preparation time and capacity needs. Data pressure likewise helps in diminishing stockpiling requests. Utilization of instance selection algorithms that attention on choosing delegate instances is normal in the event of instance-based or languid classifiers.

These kinds of classifiers don't manufacture classification models, yet store all the preparation instances. At the point when an inconspicuous instance is to be grouped, the concealed instance is contrasted and put away instances and it is characterized to a class of an instance that is nearest to the concealed instance. As the classifier needs stockpiling of all preparation instances and every concealed instance is contrasted and these put away instances, number of preparing instances decides both the time required for arranging inconspicuous instances and capacity prerequisites of the instance-based classifiers. A few instance-selection algorithms have been created and are regularly utilized in the event of instance-based classifiers.

In its easiest structure, an Artificial Neural Network (ANN) is an impersonation of the human mind. A characteristic cerebrum can adapt new things, adjust to new and evolving condition. The mind has the most astounding ability to investigate deficient and hazy, fluffy data, and make its own judgments among it. For instance other's penmanship can be perused however the manner in which they compose might be totally not quite the same as the manner in which it composed. A youngster can recognize that the state of a ball as well as orange in a circle. Indeed, even a couple of days old infant can perceive its mom from the touch, voice and smell.

This can distinguish a realized individual even from a hazy photo. Cerebrum is a very intricate organ which controls the whole body. The mind of even the most crude creature has more capacity compared to most exceptional PC. Its capacity isn't simply controlling the physical pieces of the body, yet in addition of progressively complex exercises includes reasoning, learning and so forth, exercises that can't be depicted in physical terms. An artificial reasoning machine is still past the limit of the more exceptional supercomputers.

Classification is one of the usually utilized data mining assignments. It is a managed learning procedure. Starting examination in the field concentrated on improvement of a few unique procedures to fabricate classification models, for example, neural network, instance-based classifiers and choice tree. Classification exactness, preparing time, stockpiling necessity and conceivability of the model are a portion of the metrics utilized for contrasting the exhibition of different procedures.

Neural Network (NN) classifiers give high classification precision; better speculation capacity and power to change however require high preparing time just as extra room and need intelligibility and steady learning capacity. Data decrease system was the mainstream zone of research data mining fields during the most recent decade. Advents in data decrease methods brought about the improvement of a few data decrease algorithms that can perform characteristic and instances selection. Trait selection algorithms would be ordinarily utilized for diminishing the dimensionalities of data that help in lessening the time required to prepare and space multifaceted nature of the classifiers.

II. LITERATURE SURVEY

Raju P. S. [1] Data mining and Customer Relationship Management is required in Banking and Retail Industry. This paper incorporates different errands and utilizations of data mining valuable in these businesses. The bank and retail industry understands that data mining is valuable procedure for basic leadership and gives points of interest in focused condition performed

Vidhate D. R. [2] this paper focuses on various perceptions of specialists for the advancement of super bazaar with customer reaction. In this, learning mining is utilized to break down client purchasing conduct in super bazar.

Lalithdevi B. [3] this paper clarifies about Data Mining on Web log data. In this, exercises of web utilization mining and advancements utilized in each undertaking clarify in detail. Data readiness and example age techniques are explained. That is helpful in discovering route designs. Finally presumes that many mining algorithms utilize the successive example age strategy and rest utilizes specially appointed techniques.

Bhaise R. B. [4] Education Data Mining is the principle subject of this paper, for better under study's training. For this, the creator utilized K-Means or Clustering techniques on the example data. This procedure is utilized to break down the data from various measurements and order the data. They made groups as indicated by the understudy's presentation in the examination. The data produced in the wake of executing mining strategy is particularly helpful for instructor just as understudy.

Borkar S, Rjeswari K. [5] Association rule mining is valuable to assess understudy execution in the investigation. In this paper, for data investigation Weka apparatus is utilized. The principle objective of this paper is anticipating understudy execution in the college test on premise of the criteria inner test, task, participation, and so forth. This paper inferred that the aftereffect of college result will be improved of the poor understudies by giving additional endeavors in their unit test, participation, task and graduation.

[

Chaurasia V, Pal S. (2013) [6] this paper gives the overview data of various data mining techniques for therapeutic individual for basic leadership. From this, the specialists can foresee the nearness of coronary illness. This paper utilized Naive Bayes, J48 Decision Tree, Bagging techniques in the field of coronary illness finding. Thus, the stowing calculation is better from others since it gives intelligible classification rules.

HyupRoh [7] presents half and half models with neural networks and time arrangement model for determining the instability of stock cost list in two vision focuses: deviation and heading and the outcomes demonstrated that ANN-time arrangement models can build the prescient power for the point of view of deviation and bearing exactness. These exploration exploratory outcomes demonstrated that the introduced crossover model could be improved in gauging volatilities of stock value record time arrangement.

Kim [8] introduced a hereditary calculation based instance selection technique in Artificial Neural Networks (ANN) for monetary anticipating. The transformative instance selection is utilized to decrease the dimensionality of data and may likewise evacuate the unsafe and excess instances. Likewise, transformative pursuit procedure is additionally used to locate the perfect association loads between layers in ANN. In spite of the fact that most of learning frameworks recently planned generally expect that preparation sets are well-adjusted, this supposition that isn't really right. Without a doubt, there exist numerous areas for which one class is spoken to by countless precedents while the other is spoken to by just a couple. On the off chance that 99% of the data are from one class, for most reasonable issues a learning calculation will be unable to show improvement over the 99% exactness reachable by the minor classifier that marks everything with the larger part class.

Cano [9] introduced to an instance selection strategy dependent on developmental calculation, which is a versatile technique that originates from regular advancement and exceptionally helpful for inquiry and improvement. Through various experimental examination, it is appeared with developmental calculation, better instance decrease and higher classification exactness can be effectively accomplished. Moreover, the creators called attention to that transformative calculation would be conspicuous and compelling device in the field of the instance selection technique. This is very reliable with decision that the hereditary calculation based instance selection introduce in this paper is a powerful and widespread instance selection strategy.

Japkowicz [10] talked about the impact of irregularity in a dataset. The creator assessed two resampling techniques. Irregular re-sampling comprised of re-sampling the small class indiscriminately until it comprised of the same number of tests as the greater part class, which is additionally called oversampling. Irregular under-testing was additionally considered, which required under-examining the larger part class tests until their numbers coordinated the quantity of minority class tests. She noticed that both the examining approaches were successful, and additionally seen that utilizing the refined testing techniques did not give any unmistakable bit of leeway in the area considered. Neural networks have been effectively utilized in a few broadened applications. Regardless of their higher speculation capacity, vigor to clamor and exactness they are not viably utilized in data mining. A portion of the significant pundits on utilization of neural networks for settling data mining classification undertakings would be higher preparing time, absence of intelligibility and absence of steady learning capacity. A few research endeavors have been accounted in the writing to conquer the issue of fashionable.

III. DATA MINING USING SUPERVISED INSTANCE SELECTION (SIS) FOR BETTER CLASSIFICATION

The below figure (1) shows the flow chart of supervise instance selection algorithm for better classification is done. In this initially, input data is trained. Next data is preprocessed using data pre processing block and optional features are applied to the preprocessed data. For the featured data network training is applied. Along with that testing is performed for the trained data.

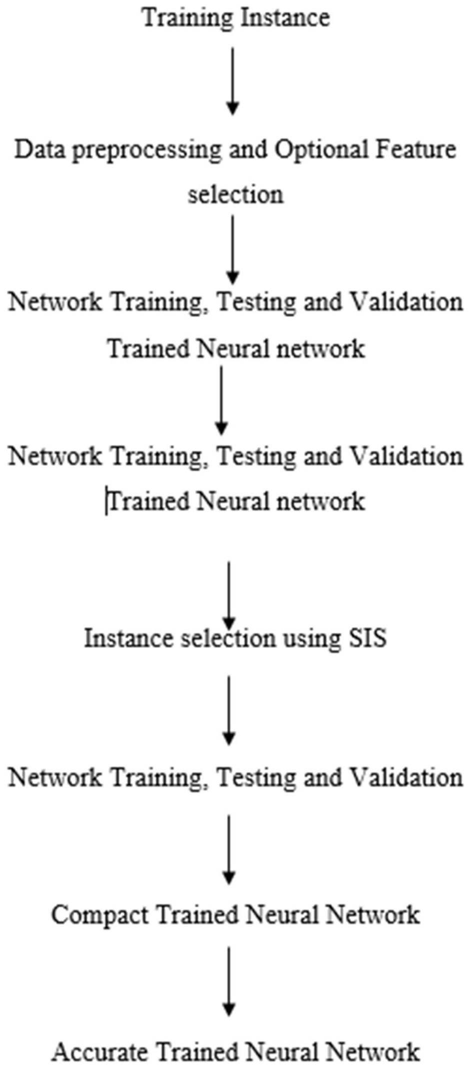
Training Instance
Data preprocessing and Optional Feature selection
Network Training, Testing and Validation
Trained Neural network
Network Training, Testing and Validation
Trained Neural network
Instance selection using SIS
Network Training, Testing and Validation
Compact Trained Neural Network
Accurate Trained Neural Network

Fig. 1: Flow Chart of Supervised Instance Selection (SIS) For Better Classification

After testing data is validated using validation trained neural network. Now, instance selection using SIS process is performed for the validate data. After this SIS process again data is trained, tested and validated. Compact trained neural network will saved the validated data and train regarding to neural network concept. At last accuracy is improved this trained neural network.

ALGORITHM:

Step. 1: In this initially, input data is trained.

Step. 2: Next data is preprocessed using data pre processing block and optional features are applied to the preprocessed data.

Step. 3: For the featured data network training is applied. Along with that testing is performed for the trained data. After testing data is validated using validation trained neural network.

Step. 4: Now, instance selection using SIS process is performed for the validate data.

Step. 5: After this SIS process again data is trained, tested and validated.

Step. 6: Compact trained neural network will saved the validated data and train regarding to neural network concept.

Step. 7: At last accuracy is improved this trained neural network.

IV. RESULTS

Accuracy: It is defined as the ratio of correctly classified instances to the total predictions predicted by classifier and is expressed as

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Precision: The precision is defined as the ratio of number of instances which are classified as True Positives (TP) to the number of instances that are classified as False positive (FP+TP).

$$Precision = \frac{TP}{(TP+FP)}$$

Recall: It is also known as sensitivity, TPR (True positive Rate). It is a measure that gives the ratio of true positives. It is defined as the ratio of number of instances which are classified as TPs to the instances that are actually positive (i.e., TP + FP).

$$Recall = \frac{TP}{(TP+FN)}$$

F1-Score: This is a weighted average of precision and recall. For a better classification performance, the must be 1 and for bad performance, it must be zero.

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

The below table (1) shows the comparison table of Semi-supervised Classification Method and Supervised instance selection (SIS). In this accuracy, precision, recall and F1 score parameters are used. Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy, precision, recall and F1 score in effective way.

Table. 1: Comparison Table

S.No	**Parameters**	**Semi-supervised Classification Method**	**Supervised instance selection (SIS)**
1	Accuracy	62%	96%
2	Precision	45%	55%
3	Recall	56%	91%
4	F1-Score	46%	54%

The below figure (2) shows the comparison of accuracy, recall for Semi-supervised Classification Method and Supervised instance selection (SIS). Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.

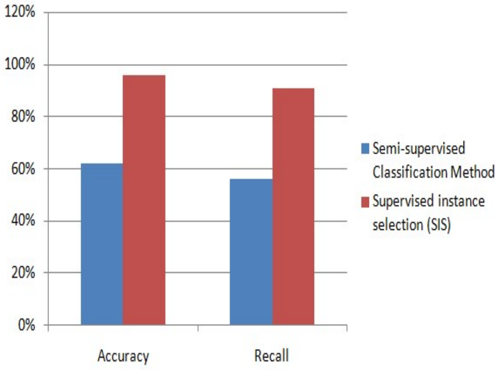

Fig. 2: Comparison of Accuracy and Recall

The below figure (3) shows the comparison of precision for Semi-supervised Classification Method and Supervised instance selection (SIS). Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.

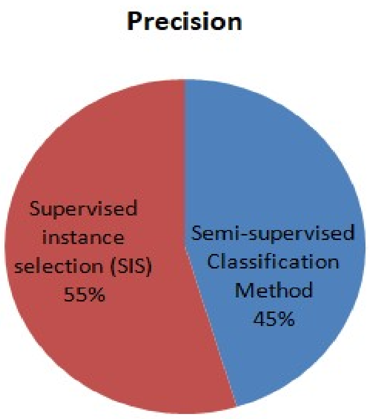

Fig. 3: Comparison of Precision

The below figure (4) shows the comparison of F1-Score for Semi-supervised Classification Method and Supervised instance selection (SIS). Compared with Semi-supervised Classification Method, Supervised instance selection (SIS) improves the accuracy and recall in effective way.

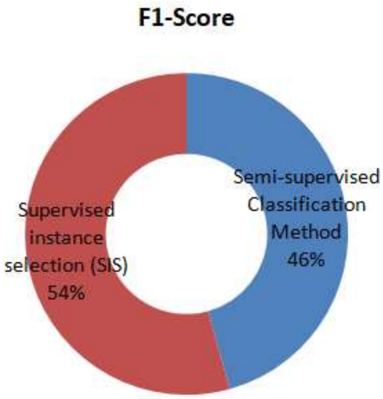

Fig. 4: Comparison Of F1-Score

V. CONCLUSION

In recent times, the instance selection has becoming increasing because of the vast volumes of data which is produced constantly in various research fields. Hence in this paper, data mining using Supervised Instance Selection (SIS) for better classification accuracy in artificial neural networks gives effective outcome. In this a new instance selection technique is presented based on the original labeled data. From results it can observe that it will improve the accuracy, Precision, Recall and F1 score compared with semi-supervised classification method.

REFERENCES

[1] Raju, P.S., Bai, V.R. &Chaitanya, G.K., 2014. Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International

Journal of Innovative Research in Computer and Communication Engineering, 2(1), pp.2650–2657.

[2] Vidhate D. R.(2014), "A conceptual study of Consumer Behavior Analysis in Super Bazar using Knowledge Mining", Sinhgad Institute of Management and Computer Application, Pages : 70-75, ISBN : 978-81-927230-0-6.

[3] Lalithdevi B., Ida A. M., Breen W. A. (2013),"A New Approach for improving World Wide Web Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 1, Pages : 243-251, ISSN : 2277 128X.

[4] Bhaise R. B. "An algorithm for a selective nearest neighbor decision rule", IEEE Transactions on Information Theory, Vol. 21, No. 6, pp.665–669.

[5] Borkar S., Rjeswari K. (2013),"Predicting Students Academic Performance Using Education Data Mining", International Journal of Computer Science and Mobile Computing, Volume2, Issue7, Pages:273-279, ISSN : 2320-088X.

[6] Chaurasia V., Pal S. (2013), "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology, Volume 2, Issue 4, Pages : 56-66, ISSN : 2296-1739.

[7] HyupRoh, "A compact and accurate model for classification", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 6, pp.203–242 2007.

[8] Kim, "Using neural networks for data mining", Future Generation Computer Systems, Vol. 13, Nos. 2–3, pp.211–229, 2006.

[9] Cano, "A general neural framework for classification rule mining", Int. J. Computers, Systems and Signals, 2003 Vol. 1, No. 2, pp.154–168.

[10] Japkowicz , "Symbolic interpretation of artificial neural networks", 2000 IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 3, pp.448–463.

[11] S. García, J. Derrac, J. R. Cano and F. Herrera, "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 417-435.

[12] Craven, M. and Shalvik, J. (1997) 'Using neural networks for data mining', Future Generation Computer Systems, Vol. 13, Nos. 2–3, pp.211–229.