2010 2nd International Conference on Education Technology and Computer (ICETC)

A General Decision Layer Text Classification Fusion Model

Xiao-Dan Zhang
Institute of Scientific and Technical Information of China, Beijing 100038, P.R. China
e-mail:zhangxd@istic.ac.cn

***Abstract*—An general decision layer text classification fusion model for higher precision, is proposed, which based on model theory of information fusion, and different classification algorithm of the feature layer fusion centre having different pre-processing, their classification results input into the decision layer fusion centre separately. And the final classification result output from decision layer fusion centre. KNN, SVM and BP Net are used in feature layer, and D-S Theory is used in decision layer. The model is realized in the experiment. From the experiment and contrast, the text classification fusion model can improve the classification precision effectively.**

Keywords- text classification; decision layer classification fusion model; classification algorithm; information fusion

I. INTRODUCTION

Text classification is the important study topic of information retrievals field, which treats and organizes large-scale of text data, and is studied and concerned by people. Though there are many sophisticated classification algorithms, such as KNN, SVM, and Bayes, how to improve the classification precision is the main hot problem.

Based on the model theory of information fusion, a decision layer text classification fusion model is proposed in this paper. The fusion model has two layers, one is feature layer, which realizes feature classification, and the other is decision layer which treats the results input by feature layer and gets the final classification result. The model has been realized in the text classification system of some information resource management system, and the experiment proves that the fusion model is effective and can improve classification precision.

II. DECISION LAYER FUSION MODEL OF TEXT CLASSIFICATION

Information fusion is the assessment process of managing more kind and type information data to get higher precision result. There are three kinds of fusion layer, which are data layer, feature layer, and decision layer, and two kinds of structures, which are series and parallel connection structures. Every layer and connection has different function [1]. Text classification can be seen as an assessment problem, which is to classify a text into a true and given class. So, fusion theory can be used to building classification fusion model for higher precision.

The model is showed as Fig.1.

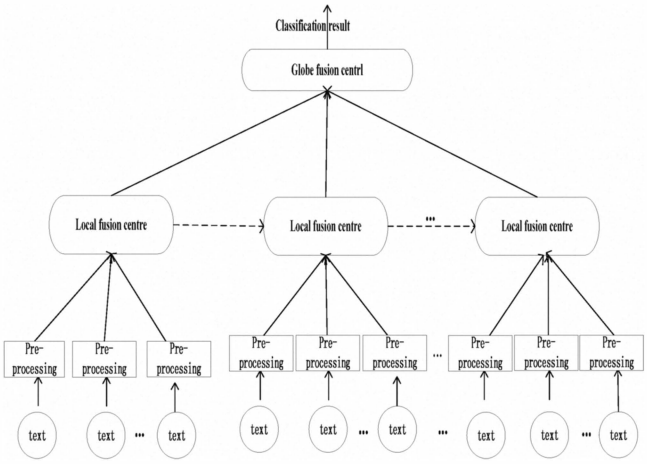

Fig.1. Decision layer fusion model

From the figure 1 we can see the fusion model has two layers, which are feature layer named local fusion center and decision layer named global fusion center. The local fusion center realized initial classification by single classifiers, every classification result input to the global fusion centre and the relational local center. The global center realized final classification by decision fusion algorithm. And the final classification result is got by the fusion model.

The specific process includes the feature layer classification and decision layer classification. Firstly, text is pre-processed, including segmentation, feature extraction and vector express. Secondly, the processed text is input into every classification of feature layer. Thirdly, the text is processed by every classification; every classification result is input into the global fusion center and the relational local center. At last, the final classification result is got by the global fusion center.

III. DECISION LAYER FUSION ALGORITHM OF TEXT CLASSIFICATION

For selecting feature layer classification algorithm, we contrast KNN, SVM, Bayes and BP Net for precision under the same experiment condition [2]. From the experiment, KNN, SVM and BP Net are as the feature layer algorithms. D-S Theory as the common decision fusion algorithm is selected as decision layer fusion algorithm in the fusion model.

The classification fusion model is showed in Fig.2.

978-1-4244-6370-1/$26.00 © 2010 IEEE

Authorized licensed use limited to: Prince of Songkla University provided by UniNet. Downloaded on July 15,2021 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

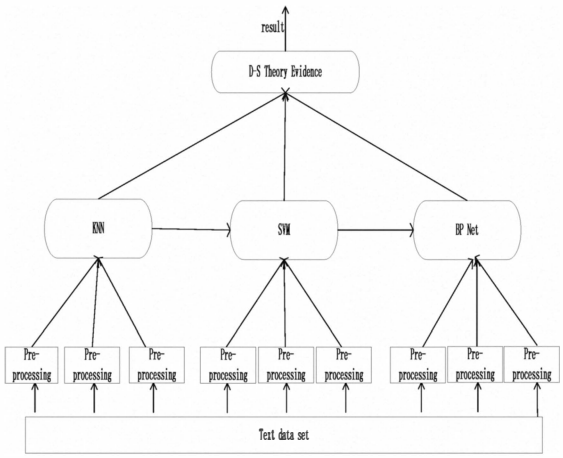

Fig.2. Process of decision layer fusion algorithm

From Fig. 2. we can see, KNN, SVM, and BP Net adopted as feature layer fusion algorithm and D-S Theory algorithm used in decision layer.

The steps of the classification fusion algorithm are as followed:

Step 1, Training texts are input into KNN, SVM and BP Net, and the KNN classifier model, SVM classifier and Bayes classifier are determined.

Step 2, the classification results of KNN, SVM and BP Net are input into voting algorithm, the final classification result is got.

Step 3, Test texts are input into KNN,SVM and BP Net, the results of the three kinds classification model are input into decision layer , that is D-S Theory Algorithm, and the final result is output.

If U is all possible value set of C, and all elements are incompatible， then U is as identification framework of C.

Define 1: if U, then function m: 2U->[0， 1] meet condition : , then m(A) is the basic probability value,

Define 2: if $m_1$and $m_2$are the two basic probability of U, focal elements are $A_1, A_2,\ldots,A_k$ and $B_1, B_2,\ldots,B_k$, then,

$$m(C)=\begin{cases}\dfrac{\sum_{\substack{i,j\\ Ai\cap Bj=C}} m1(Ai)m2(Bj)}{1-K1}, \hat{a}\ C\subset U, C\neq\varnothing\\ 0, C=\varnothing\end{cases}$$

$$K1=\sum_{\substack{i,j\\ Ai\cap Bj=\varnothing}} m1(Ai)m2(Bj).$$

If m1 is the basic probability value of the main evidence focal element, evidence relation factor p, 0<p<1. Two evidences are more equal if p is bigger. When p=1, it is the basic D-S evidence fusion.

For the evidence of the same type, different feature layer output is fused by DS theory, then the final classification is got from the decision layer.

IV. EXPERIMENT

The decision layer classification fusion model is used in the Computer Center of some Department. For proving the effectiveness of the model, we adopt the same training set, testing set and classification system for KNN, SVM, Bayes and the fusion method. JAVA is used as the development plot.

The contrast results are showed as followed table 1.

TAB.1. THE CONTRAST RESULTS OF KINDS OF ALGORITHMS

Methods\ precise	Recall rate	precision	F1
KNN	80.3%	90.2%	86.3%
SVM	76.4%	92.6%	87.6%
BP Net	74.1%	89.6%	84.9%
Fusion model	85.6%	94.2%	89.7%

In Tab.1, we can see the precision and recall rate of fusion method better than the other methods.

From Tab.1 we can see that the precision of the new fusion method proposed in the paper is exceeded to the old classification method.

V. CONCLUSION

A new decision classification fusion model and algorithm is proposed in this paper. That is the classification results of the local fusion center are input into global fusion center, then the final classification result is got from decision layer fusion algorithm. The process is that texts are input into the local fusion center, and their results are input into decision layer. We adopt KNN, SVM and BP Net as feature fusion algorithm, and D-S Theory algorithm as decision layer. For proof the effectiveness of the model, we adopt the same training set and test set for KNN, SVM, BP Net and the fusion model. The contrast result proves that the fusion method is exceeded to the other methods in precision. And the fusion method is used in the Computer Center of some Department.

ACKNOWLEDGMENT

The author acknowledges the support of the Natural Science Foundation of P. R. China.

REFERENCES

[1] Zhang, B., Chen, Y., Fan, W., Fox, E. A., Goncalves, M., Cristo, M. & Calado, P.(2005a). Intelligent GP fusion from multiple sources for text classification. InProceedings of the 14th ACM international conference on Information and knowledgemanagement (pp. pp. 477-484). : ACM Press, Bremen, Germany

[2] Zhang, XD.. Text classification based on decision fusion model. In Proceedings of the 28th annual international ACM SIGIRconference on Research and development in information retrieval (pp. pp. 266-273). :ACM Press, Salvador, Brazil

[3] Zhang, G. P. (2007). Avoiding Pitfalls in Neural Network Research. Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 37, pp. 3-16.

Authorized licensed use limited to: Prince of Songkla University provided by UniNet. Downloaded on July 15,2021 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.

[4] Zhang, L., Zhu, J. & Yao, T. (2004). An evaluation of st..atistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3,pp. 243-269.

[5] Zhang, Y., Zincir-Heywood, N. & Milios, E. (2005c). Narrative text classification for automatic key phrase extraction in web document corpora. In Proceedings of the 7th annual ACM international workshop on Web information and data management (pp. pp.51-58). : ACM Press, Bremen, Germany

[6] Zheng, Z. & Webb, G.I. (2000). Lazy Learning of Bayesian Rules. Machine Learning, 41, pp. 53-84.

[7] Xu L Y, Du Q D. Application of neural fusion to accident forecast in hydropower station. Proceedings of the Second International Conference on Information Fusion. Vol2 Sunnyvale, 1999, pp. 1166-1171.

[8] Schapire, R. E., Singer, Y. & Singhal, A. (1998). Boosting and Rocchio applied to text filtering. In Proceedings of SIGIR-98 21st ACM International Conference on Research and Development in Information Retrieval (pp. p. 215--223). : ACM Press New York US

[9] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34, pp. 1-47.

[10] Shih, L. K. & Karger, D.R. (2004). Using urls and table layout for web classification tasks. In Proceedings of the 13th international conference on World Wide Web (pp. pp.193-202). : ACM Press, New York, NY, USA

[11] Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (pp. pp. 41-46). : T.J. Watson Research Centre, Seattle, Washington

[12] Riloff, E. (1995). Little words can make a big difference for text classification. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. pp. 130-136). : ACM Press, Seattle, Washington, United States

Authorized licensed use limited to: Prince of Songkla University provided by UniNet. Downloaded on July 15,2021 at 09:40:42 UTC from IEEE Xplore. Restrictions apply.