Bulletin de la Société Royale des Sciences de Liège, Vol. 85, 2016, p. 358 - 375

A New Synthetic Oversampling Method Using Ontology and Feature Selection in Order to Improve Imbalanced Textual Data Classification in Persian Texts

Jafar POURAMINI[1], Behrouz Minaei-BIDGOLI[2]

[1] Department of Information Technology, Faculty of Engineering, University of Qom

j_pouramini@pnu.ac.ir

[2] Faculty of Computer Engineering, Iran University of Science and Technology

b_minaei@iust.ac.ir

Abstract

Ever-growing extension of textual data has increased the necessity of processing textual data. Data imbalance in classification of textual data is one of the cases that decrease efficiency. In order to confront with imbalance problem, various methods are suggested. Some of the methods are: data-based, cost-based, algorithm-based and feature selection methods. In recent researches, some methods are considered into account using ensemble methods. In this research, a new oversampling method is suggested. In the new method the number of minor class samples is increased using ontology and then random oversampling is performed for minor class. Finally, using the methods of feature selection, appropriate features are selected. New ensemble method was tested using Hamshahri data. The results show that the ensemble method on Hamshahri collection, despite decreasing number of features, causes the improvement of classification results for polynomial Naïve Bayes and decision tree.

Keyword: Oversampling, Ontology, Imbalanced, Feature selection.

1. Introduction

By increasing data, analysis and discovering knowledge of raw data is turned to a big problem. Though available knowledge discovery methods and engineering techniques of data have shown big success in many practical programs, but the problem of learning from imbalanced data is a big challenge which has attracted university and industry to itself. Imbalanced data is said to a collection of data that the samples of one class are much lower than

the samples of other class. The class with high number of samples is called major class and the class with very low number of samples is called minor class. In some cases, the nature of data is imbalanced; this kind of imbalance is called intrinsic. For example, data of the areas like cheating discovery, cancer diagnostic, earthquake forecast and texts' classification are imbalanced. But in some cases, because of the limitations like high cost of collecting samples, legal problems and/or private issues, the possibility of data gathering is not sufficient, and this causes imbalance in data; this state of imbalance is called extrinsic [1]. Classification algorithms have higher tendency to major class and even may collide with data of minor class as deviated data. Imbalanced data are occurred in many contexts. Textual data are one of the areas that imbalanced is occurred. For example, when a collection with diverse subjects is divided to two classes such that one class includes documents of a specific subject and another class includes documents with other subjects. This situation is called one-against-all [2].

Learning from imbalanced dataset needs principles, algorithms and tools for efficiently converting raw data to information and presenting knowledge. In order to resolve the problems related to imbalanced textual datasets, various approaches are introduced which are: data level, cost-sensitive learning framework, feature selection, term weighting approach, ensemble approach [2,3]. By investigating the published papers in valid magazines, the researches that are perform in this regard can be divided into two classes based on approach. These two approaches are:

Statistical approach: In this approach, statistical features are intended; the parameters like number of words' repetition in one document, number of word-contained documents, word repetition in one class and distributing word in different documents are used. In this approach, some special features of text like arrangement of words, relations between words and words' dependency are neglected. In this approach, pre-processing of removing words is usually stopped and rooting is performed.

Linguistic approach: In this approach, specific features of text are used. Arrangement of words, meaning of words, word's dependency to each other, relation between implications and role of words in sentence are used. In this approach, it is possible to stop pre-processing of removing words and perform rooting. The references out of texts like ontology may be also used. Yanling et al (2010) investigated imbalanced textual data [4]. They investigated effective factors

on efficiency of imbalanced textual data classification. Yanling et al announced effective factors as following:

Data distribution: The ration of samples with minor class to major class is one of the important factors in efficiency of classifying imbalanced data [4]. This ratio is different in different applications. According Joshi investigations, this ratio is equal to 1:10 in some applications while it is equal to 1:35 in some other applications and when the ratio is exceeded from this limit, classification efficiency is considerably decreased [5].

Classes' overlap: if there is too much overlap between classes, imbalanced distribution and efficiency of classifiers will be more decreased [6]. But by the increase of classes' overlap, linear classifiers' sensitivity to data distribution increases [7].

Educational data size: If the ratio of minor and major classes are assumed constant, data size can affect efficiency of classification such that whatever the number of samples is lower, efficiency is decreased more [8].

Existence of subclasses: existence of sub-classes causes the complexity of education in classifiers [9].

According these researches, efficiency is decreased when number of available samples is very low for minor class. Persian language is a conventional language in Middle East. Iran, Afghanistan and Tajikistan are among the countries that Persian is common in them. Persian writing is from left to right similar to Arabic. Persian has 32 alphabets and is similar to Arabic.

According these researches, efficiency is decreased when the number of available samples for minor class is very low. In this research, a new ensemble method is presented for oversampling and feature selection for imbalanced textual data with low number of samples in Persian texts. Different sections of this writing are as following: in second section, the performed works are investigated. In third section, the suggested method is introduced. In fourth section, performed tests are presented. In fifth section, the results are discussed and investigated and in sixth section, the results are presented.

2. Related work

Data-based methods are trying to change data imbalance via increasing minor class samples and/or decreasing majority samples. But these methods have the problem of side effects [10]. Oversampling methods that only increase minor class cause overfitting [11] and the methods that

cause the decrease of major class samples cause loss of useful data in major class. Therefore, researchers have presented some methods to overcome this problem in different data area. Chawla [12] presented SMOTE method for decreasing imbalanced effects of data. This method causes the increase of samples in minor class by adding new samples between existing samples. This method effectively avoids from overfitting. Simul [13] presented MWMOTE. In this method, the samples that are hardly learned by algorithm are specified; then a weight is considered for these samples. This weight is determined in terms of Euclidean distance between sample and its nearest sample from major class. Sun et al [14] performed many studies and tests on classifying imbalanced texts using support vector machine classifier. In this research, sampling methods like sub-sampling and pre-sampling were investigated and the results were compared by the other method of collision with imbalanced data like the method of using cost for classification error.

Enhong et al [10] presented a new method named DCOM for producing new samples of minor class to improve imbalanced text classification. In this research, a probabilistic topic model was created for minor class using overall semantic information of minor class and then new samples are produced using it. The benefit of the presented method is the reduction in overfitting probability.

Borajo et al [11] using Hidden Markov model presented a new oversampling method for producing new documents named COS-HMM. COS-HMM model is trained using collection. Then this model is used as production engine of new ensemble documents. This new model showed better efficiency compared to SMOTE and ROS methods.

Tang et al [15] combined sampling method and feature selection method and their combination for improving imbalanced textual data classification with high dimensions. According the test results of this research, the influence of feature selection methods is more than sampling methods. The results of this research showed that applying oversampling methods before performing feature selection improves efficiency.

Zhou et al [1] used synthetic sampling. In this method, minor samples are increased using SMOTE and major samples are decreased using sub-sampling. Many classifier were created from the combination of minor class samples and all samples of major class and their combination made a classifier by ensemble method and used the result as main classifier.

3. Proposed method

The aim of this research is to present a multi-stage ensemble method and investigating its performance. In the suggested method, two methods of oversampling and then feature selection were performed. Also, in this research, text-specified oversampling method is presented. First the new oversampling method is presented.

Ontology-based oversampling method

According the performed investigations in valid scientific references, until now ontology is not used in oversampling methods of textual data. In this research, new method of oversampling is presented using ontology. For this purpose, synset of a term is determined using ontology. For determining the synset of word equivalence in each document, Word Sense Disambiguation methods are used. In disambiguation methods, the emphasis is on one document [16] and other documents having that word are not considered, and the characteristics of that term is not used in other documents in the class. Since all documents of a class are about one subject, other documents can be used for determining synset. Therefore, all documents of the class that the document exists in are used in suggested method instead of using the document with word. In order to explain the suggested method, the following relations are used.

Doc_coverage(term,doc): This criterion shows the value of word "term" in the document "doc". Whatever, this criterion is larger, it shows higher importance of the word term in the document doc. Relation 1 is used for calculating this criterion.

$$doc\_coverage(term, doc) = \frac{count(term, doc)}{length(doc)} \quad (1)$$

Count(term,doc): It is the repetition of the word term in the document doc.

Length(doc): Number of words in document doc

In order to clarify a word's value in a class, sum of doc_coverage of the intended word is calculated for all class's documents. Therefore, the result of sum is multiplied in the fraction of documents that have the word; for this purpose, relation 2 is used.

$$\text{coverage(term)} = \sum_{doc_i=0}^{n_c} doc\_coverage(term, doc_i) \quad (2)$$

n_c : represents the number of class documents

In order to determine synset, all synsets of the word are extracted. Then, sum of value for synsets' words are calculated. A synset with highest value is selected. Using relation 3, synset of each word is specified.

$$synset\ (term) = \underset{s \in synset(term)}{argmax} \sum_{w_i \in s} \text{coverage}(w_i) \qquad (3)$$

A document of minor class is selected in order to producing new sample. This document is considered as base document of producing new sample. For producing new document, each word of base document is replaced by synset words of that word which is obtained using relation 3. This causes the production of a new document which is similar to the existing document in terms of features. Because of adding some words that include all possible states of that word, it is also in minor class with regard to all documents. For example if in three different documents of minor class, the words "ماشین", "اتومبیل" and "خودرو" exist, one the words is selected with regard to that which mentioned word is most used in all documents of minor class (to sample of synset only including "خودرو"). All three new documents that are produced include the word "خودرو". As the result, dispersion is lower in three new documents and there is higher concentration on minor class words. Also, the words that don't exist in educational set, but exist in test set are considered. If a word is not found in ontology, that word is identically added.

After producing new documents, features are scored using DFS method which is one of the filter methods [17]. Relation 4 shows the method of calculating the score of word t. In this relation, t is the intended word, C_i is the class of i and M is the number of classes.

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \qquad (4)$$

$P(t|\bar{C}_i)$: The possibility of existing the word t provided the existence of classes except C_i

$P(C_i|t)$: The possibility of existing the class C_i provided the existence of word t

$P(\bar{t}|C_i)$: The possibility of lacking the word t provided the existence of the class C_i

4. Experimental results

Hamshahri standard collection [18] is used for testing the algorithms. Size of this collection is 700MB (Unicode in the form CLEF) and includes 166774 reports of Hamshahri newspaper in 82 classes from April 23, 1996 to February 11, 2003. Number of words in this collection is 417,339. Each report has averagely 380 words. This collection is one of standard collections of Persian that is approved by researchers; and until now various researches are published on it [18] [19]. Table 1 shows the names of 12 classes of Hamshahri collection. Figure 1 shows the classes and distribution of these classes [18].

Table 1 : 12 important classes of Hamshahri collection

Row	Tag	Name
1	Adabh	Art-Literature
2	Akhar	Short news
3	Ejtem	Society
4	Elmif	Science & Culture
5	Eqtes	Economy(in Iran)
6	Gungn	Miscellaneous
7	Havad	Social Event
8	Kharj	Foreign news
9	Maqal	Miscellaneous
10	Shahr	Tehran &Municipal affairs
11	Siasi	Politics
12	Vrzsh	Sport

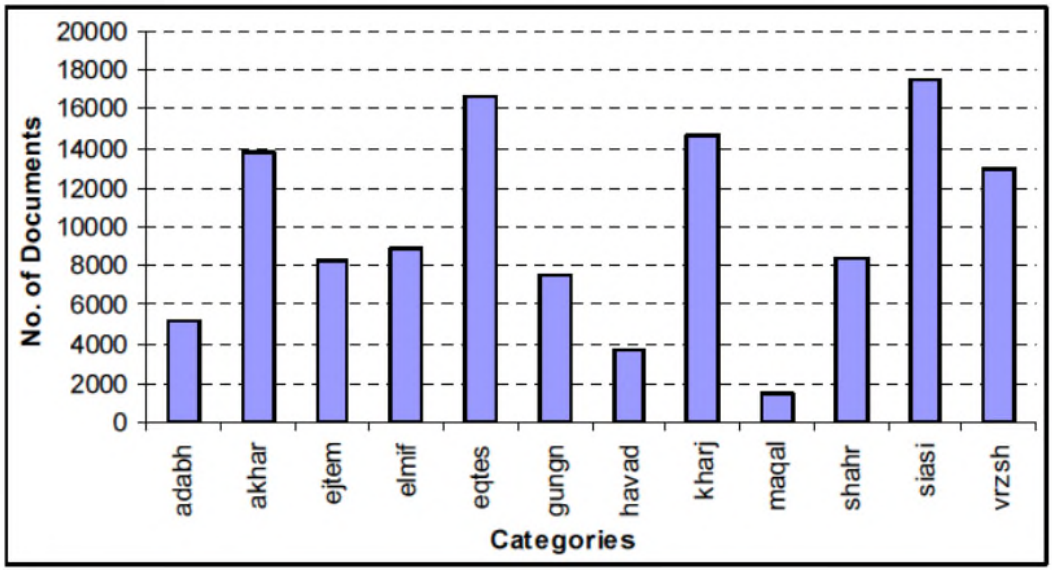

Fig. 1: Distribution of the classes of Hamshahri collection

After separating words and removing stop words, semantic rooting tool for Persian version 1.6, web technology lab of Ferdowsi University of Mashhad for rooting.

Farsnet ontology was used for tests. Farsnet is a Persian ontology which includes 10000 synset and 18000 words. Farsnet words are presented in three types of name, verb and adjective. Farsnet includes relations like Hypernyms, Synonyms, Antonyms, Meronomy [20].

Classifiers of polynomial Naïve Bayes and decision tree C4.5 were used for classification. For performing the tests k-folded cross validation was used. Data were divided to 5 sections and in each stage, 4 sections were used for education and 1 section was used for test. The average of 5 stages is considered as efficiency. In order to have more general evaluation, imbalanced rate was increased from 1 to 10% and measures were evaluated for each of rates and classifiers. Efficiency of suggested method was compared to random oversampling methods and SMOTE. For simplicity, a name was considered for each method. Table 2 shows different methods and operation that is performed in each one.

Table 2: Different methods and operation of each method

Row	Method	Explanation
1	Source	Main data without oversampling and without feature selection
2	Normal	Feature selection by DFS method (without oversampling)
3	SMOTE	Oversampling with 20% rate by SMOTE method and then feature selection by DFS method
4	ROS	Random oversampling with 50% rate and then feature selection by DFS method

5	Ontology & Normal	Oversampling using the suggested ontology method and then feature selection by DFS method
6	Ontology & SMOTE	Oversampling with the suggested ontology method and then oversampling with SMOTE method with 20% rate and finally feature selection by DFS method
7	Ontology & ROS	Oversampling with ontology method and then random oversampling with 50% rate and finally feature selection by DFS method

All measures are not appropriate for evaluating imbalanced data classifiers. Haibo compared these measures for the state of imbalanced data [21]. Different criteria exist for investigating and evaluating the results of classifiers [22]. Relations 5 to 9 show the method of calculating criteria.

$$\text{Precision} = \frac{tp}{tp + fp} \tag{5}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{6}$$

$$\text{F1\_measure} = \frac{2 * presion * recall}{precision + recall} \tag{7}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fn + fp} \tag{8}$$

ErrorRate=1-accuracy (9)

Precision is sensitive to the method of data distribution; while recall is not dependent in data distribution. As the result, one of appropriate measures in imbalanced state can be recall. F-measure used the combination of two measures of precision and recall. Relation 10 shows the way of calculating F-measure.

$$F - Measure = \frac{(1 + \beta)^2 . Recall . Precision}{\beta^2 . Recall + Precision} \tag{10}$$

In this measure β is a coefficient that shows the importance of precision measure against recall. β coefficient is determined by user.

Correctness measure is not appropriate for the state of imbalanced data [23, 24]. This criterion doesn't show algorithm operation about classification of imbalanced data. Since error measure has direct relation with correctness measure, so two measures of correctness and error are not appropriate for evaluation [25].

G-mean combines correctness rate of recognizing positive and negative samples. G-mean measure is obtained via geometrical mean of TP, TN rate. This measure is not appropriate for evaluating efficiency in imbalanced state [3, 25]. Relation 11 shows the method of calculating G-mean.

$$G-mean=\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}} \qquad (11)$$

In most cases, it is preferred for classifier to work well for minor class, even if it has low efficiency for major class [25, 26].

Graphs 2 to 9 show the efficiency of the suggested method and combination of suggested method compared to other methods in terms of precision, recall, G-Mean and F-Measure. Each graph presents the efficiency of 7 methods in table 1 in terms of one of the measures and for classifiers of decision tree and/or Naïve Bayes. Horizontal axis shows imbalanced rate which increases from 1 to 10%.

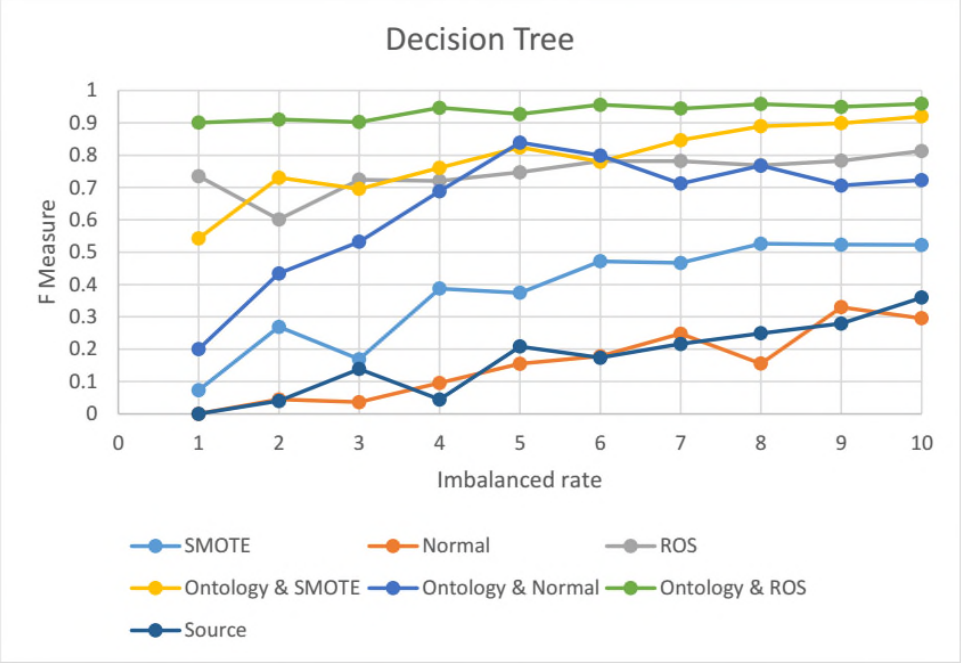

Fig. 2: Comparison of efficiency for different methods and the suggested method based on F-Measure in decision tree

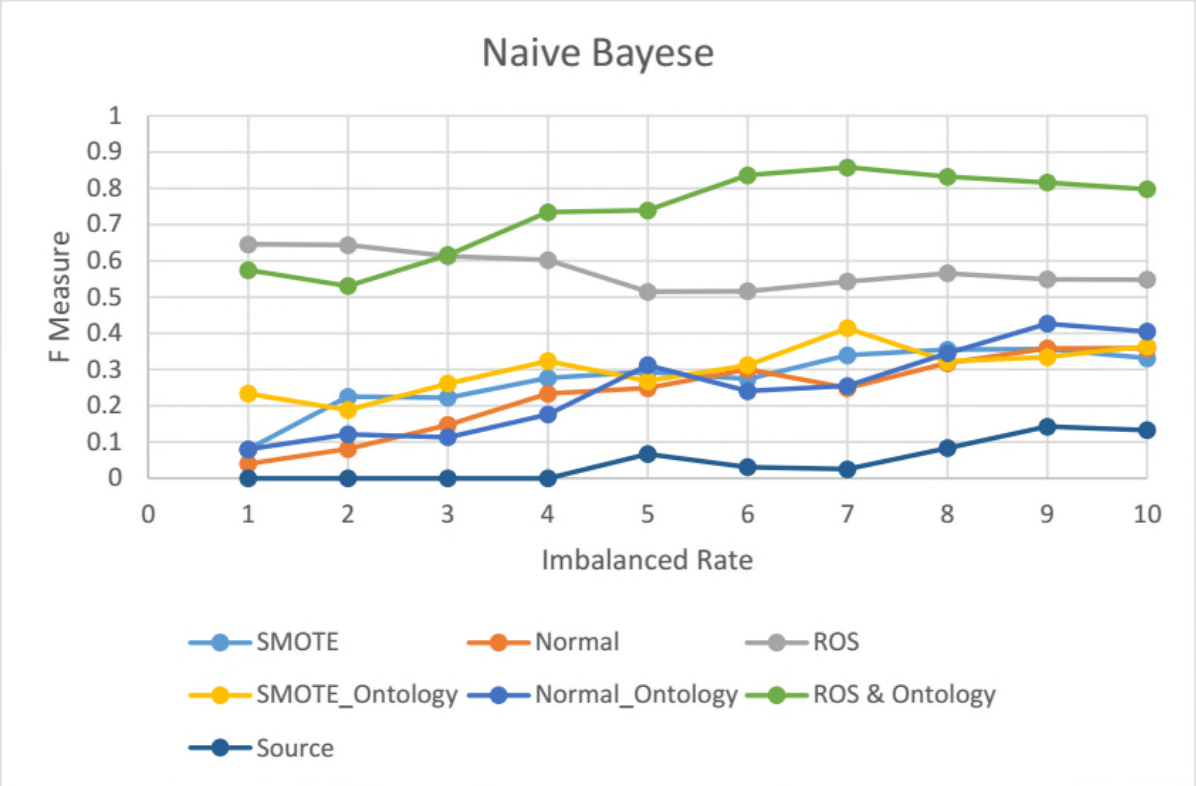

Fig. 3: Comparison of efficiency for different methods and the suggested method in terms of F-Measure in Naïve Bayes

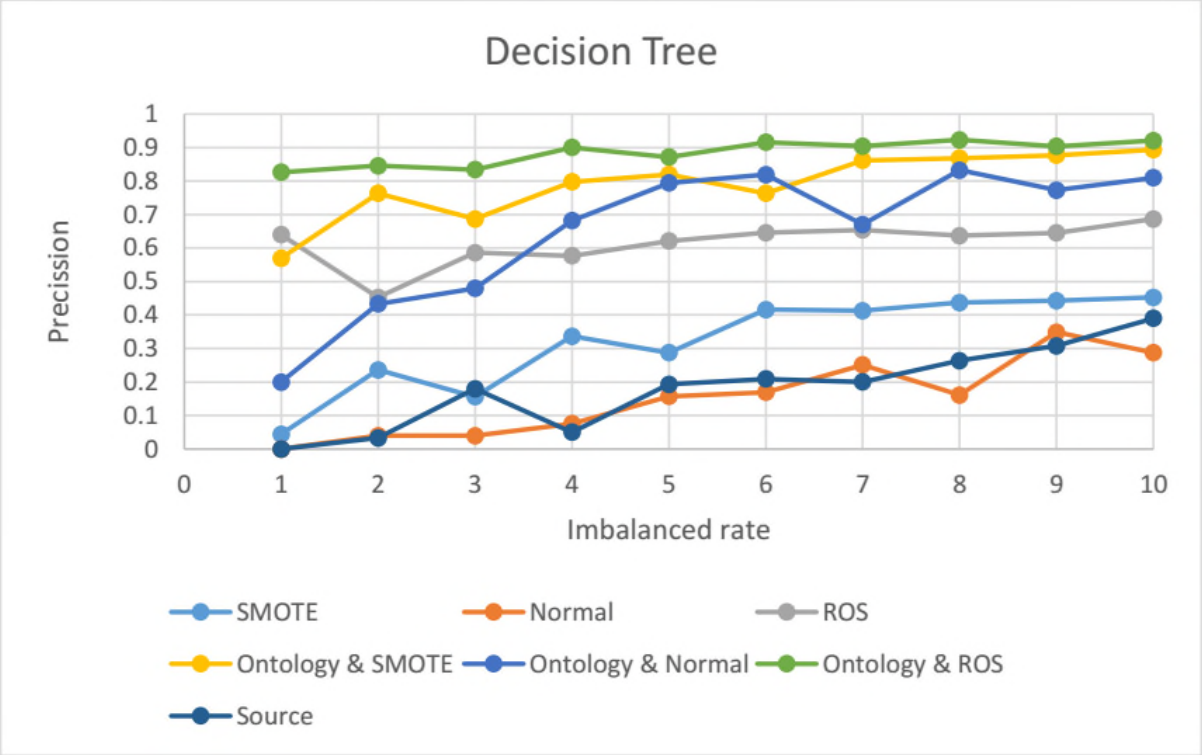

Fig. 4: Comparison of efficiency for different methods and the suggested method in terms of precision measure in decision tree

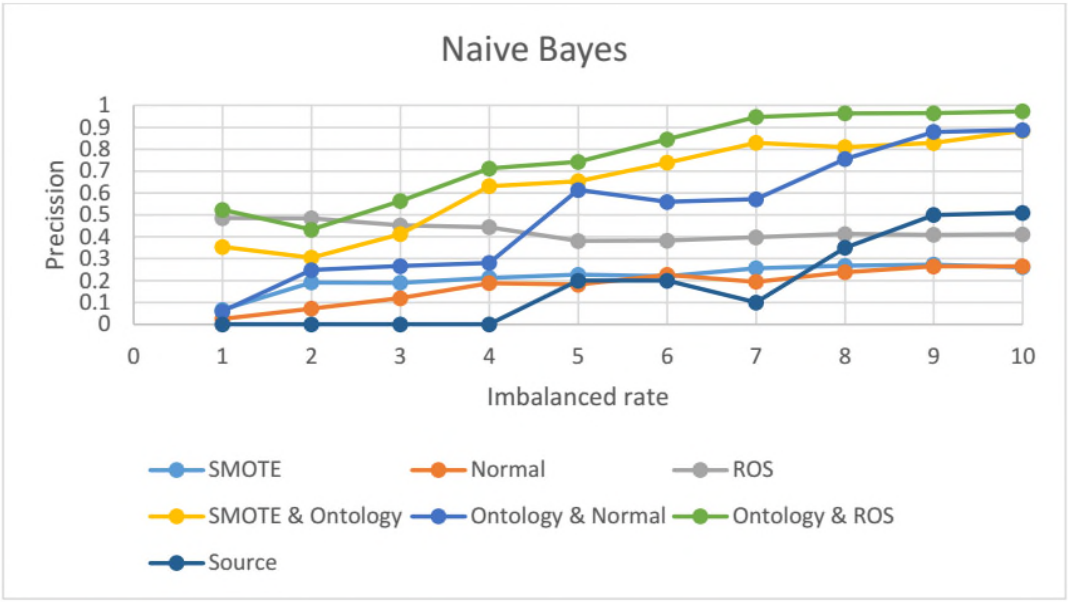

Fig. 5: Comparison of efficiency for different methods and the suggested method in terms of precision measure in decision tree

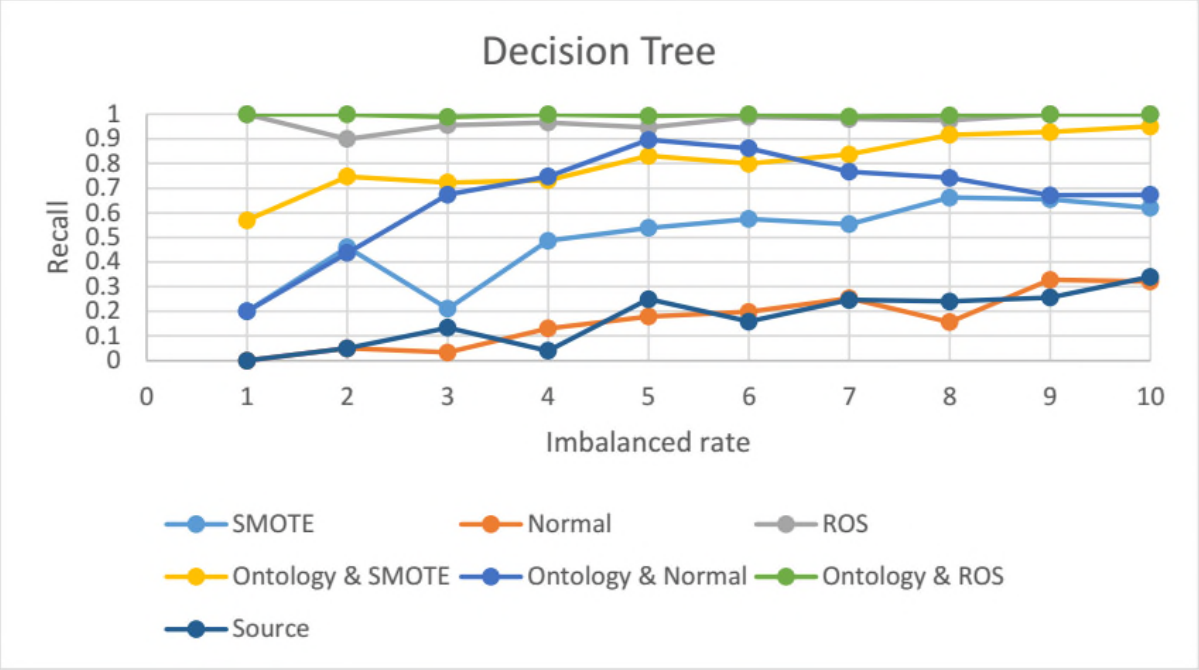

Fig. 6: Comparison of efficiency for different methods and the suggested method in terms of recall measure in decision tree

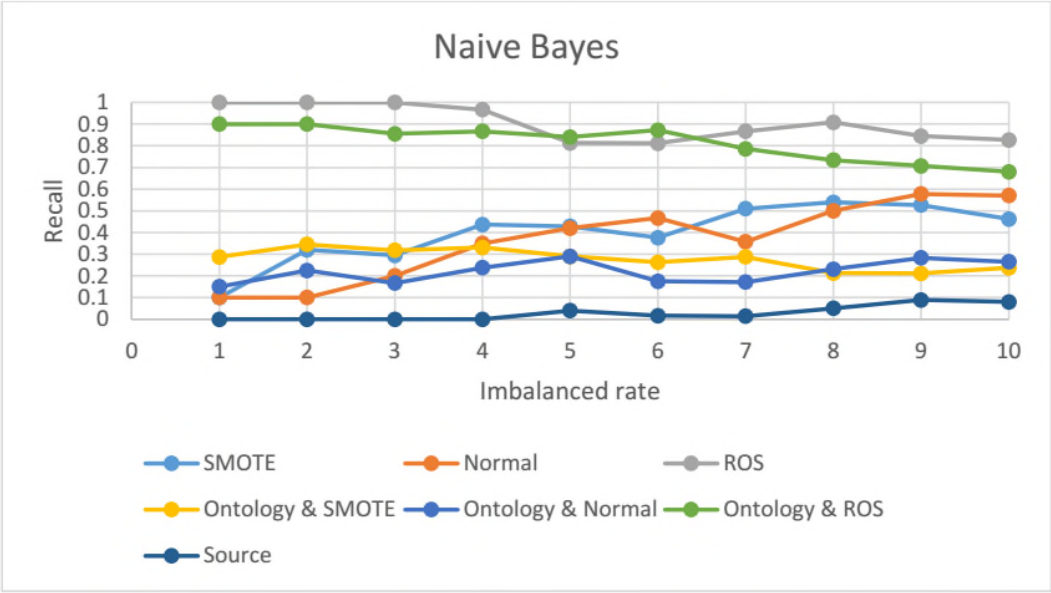

Fig. 7: Comparison of efficiency for different methods and the suggested method in terms of recall measure in Naïve Bayes

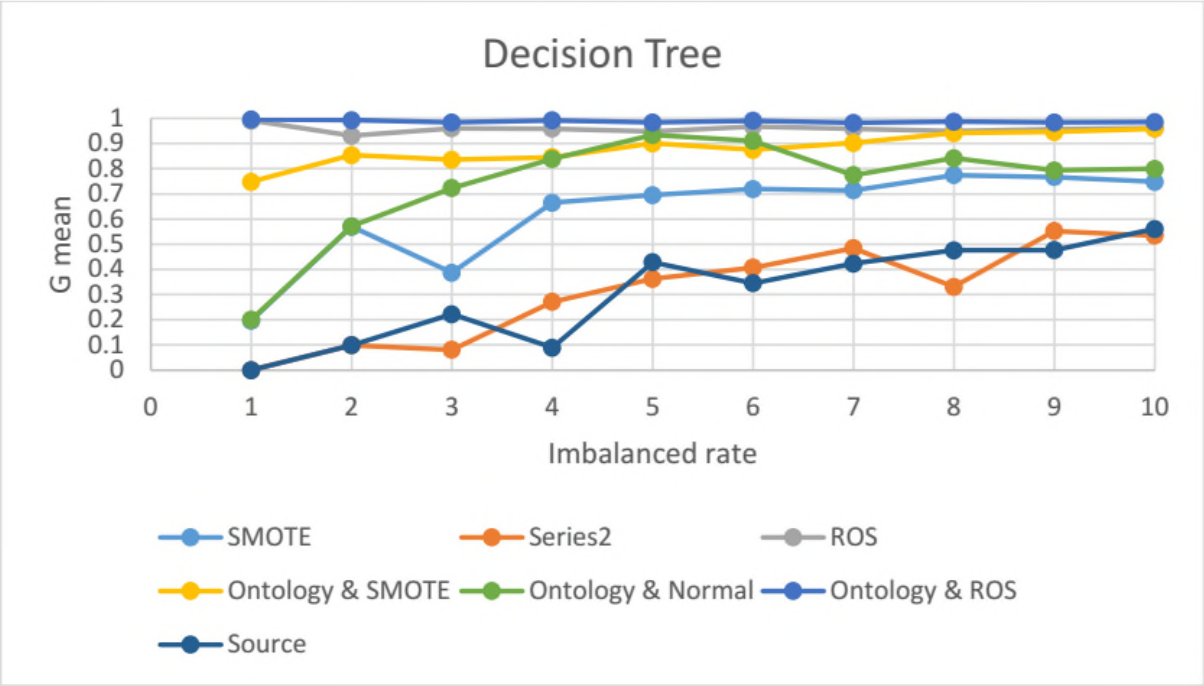

Fig. 8: Comparison of efficiency for different methods and the suggested method in terms of G-mean measure in decision tree

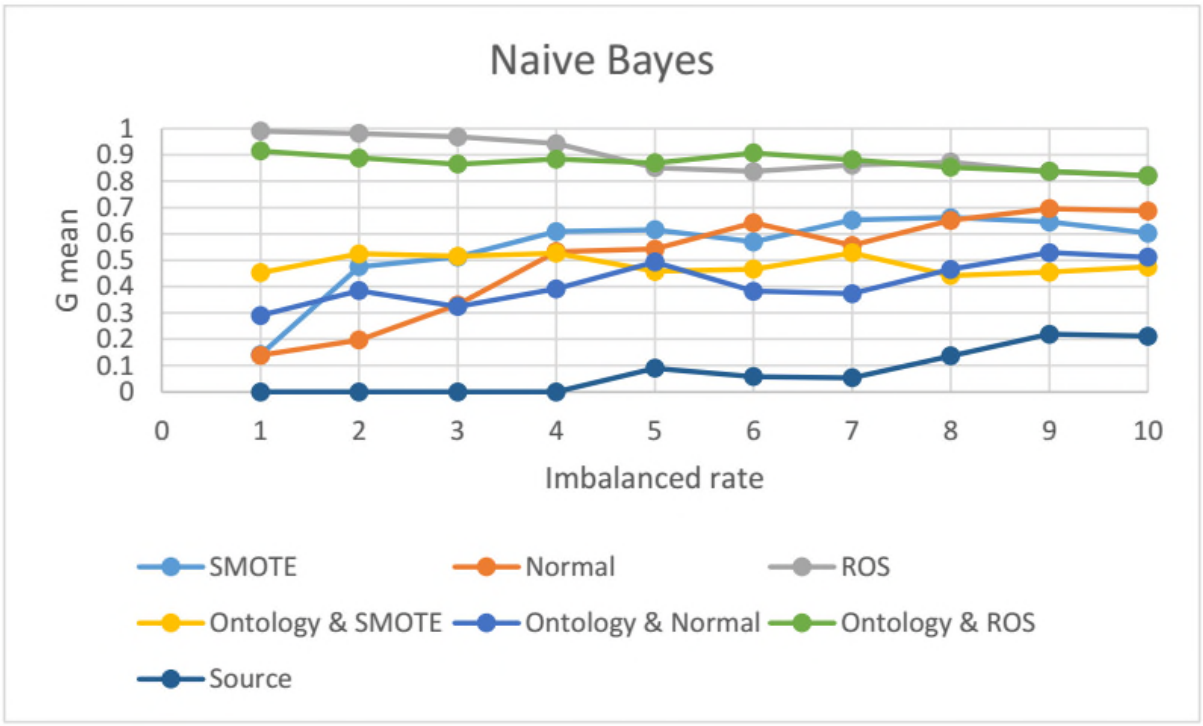

Fig. 9: Comparison of efficiency for different methods and the suggested method in terms of G-mean measure in Naïve Bayes

5. Discussion

With regard to the graphs, by comparing source method (main data without applying any method) and normal method that feature selection is merely used in it, there is no considerable difference and the superiority of using feature selection can't be observed. The reason is imbalanced data and also small number of samples. Using feature selection causes the increase of implementation speed of algorithms.

Investigation of the graphs 2 to 9 shows that the combination of suggested oversampling and random oversampling and finally feature selection in decision tree classifier has the best efficiency in F-measure, G-mean, precision and recall measures. Also, this ensemble method in Naïve Bayes causes the presentation of the best efficiency in F-measure and precision measure, and in the measures of recall and G-mean is superior of two methods.

The tests are performed for the change of imbalanced rate from 1 to 10%. Table 3 shows the mean of all measures. With regard to the table, the suggested method is superior than all measures and in G-mean and recall measures with Naïve Bayes is second one.

Table 3: Mean of measures for different imbalance rates

		Naïve Bayes				Decision Tree			
Row	Name	G mean	Recall	Precision	F measure	G mean	Recall	Precision	F measure
1	Source	0.076	0.028	0.186	0.048	0.311	0.1713	0.182	0.170
2	Normal	0.497	0.364	0.177	0.233	0.312	0.1648	0.153	0.153
3	SMOTE	0.548	0.399	0.216	0.275	0.623	0.496	0.322	0.378
4	ROS	**0.896**	**0.896**	0.426	0.573	0.958	0.971	0.614	0.745
5	Ontology & Normal	0.414	0.219	0.512	0.247	0.738	0.667	0.649	0.640
6	Ontology & SMOTE	0.484	0.278	0.644	0.302	0.880	0.803	0.790	0.7886
7	Ontology & ROS	0.872	0.814	**0.766**	**0.733**	**0.987**	**0.996**	**0.884**	**0.935**

Table 4 shows the comparison of efficiency results for ensemble method with classifying efficiency of main data. This comparison shows that efficiency shows considerable improvement than main data.

Table 4: Comparing the efficiencies of the suggested method and main data

	Naïve Bayes				Decision Tree			
Name	G mean	Recall	Precision	F measure	G mean	Recall	Precision	F measure
Source	0.076	0.028	0.186	0.048	0.311	0.171	0.182	0.170
Ontology & ROS	0.872	0.814	0.766	0.733	0.987	0.996	0.884	0.935

T Student test was used in order to investigate significance of differences in the obtained results. The following assumptions were considered.

H0: There is no significance difference between the results of Ontology&ROS method and source method.

H1: There is a significance difference between the results of Ontology&ROS method and source method.

These assumptions are also considered for other methods. These assumptions were tested using SPSS software. The results of assumption test are shown in table 5. Since in all cases $p<0.05$, so H0 assumption is rejected; therefore, there is a significant difference between the results of suggested method and other methods.

Table 5: The results of significance tests

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Ontology_ROS - SMOTE	.55695	.13771	.04355	.45843	.65546	12.789	9	.000
Ontology_ROS - Normal	.78140	.09505	.03006	.71341	.84940	25.996	9	.000
Ontology_ROS - ROS	.18977	.04691	.01483	.15622	.22333	12.792	9	.000
Ontology_ROS - Ontology_SMOTE	.14658	.09611	.03039	.07783	.21533	4.823	9	.001
Ontology_ROS - Ontology_Normal	.29521	.17864	.05649	.16742	.42300	5.226	9	.001
Ontology_ROS - Source	.85191	.05755	.01820	.81074	.89308	46.808	9	.000

6. Conclusion

In this research, a new oversampling method was presented using Persian ontology for classifying imbalanced textual data of Persian literatures. For this purpose, first equivalent synset of each word was extracted using all documents of minor class. Then, new documents were produced using these synsets. The new method was investigated singly and also by combination to other methods like ROS, SMOTE. Decision tree and Naïve Bayes classifiers were used for evaluation. The results showed that the suggested method combined with ROS methods creates highest improvement in the efficiency of classifiers. For future works, other linguistic features like concurrence of words and the role of words in sentence can be considered.

7. References

[1] P. Yang, W. Liu, B. B. Zhou, S. Chawla, and A. Y. Zomaya, "Ensemble-based wrapper methods for feature," *springer,Advances in Knowledge Discovery and Data Mining,* vol. 7818, pp. 544-555, 2013.

[2] H. Ogura, H. Amano, and M. Kondo, "Comparison of metrics for feature selection in imbalanced text classification," *Expert Systems with Applications,* vol. 38, pp. 4978-4989, 2011.

[3] S. Maldonadoa, R. Weberb, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *National Research Council of Canada, Ottawa, Canada Information Sciences,* vol. 286, pp. 228–246 2014.

[4] G. S. Yanling Li and Y. Zhu, "Data imbalance problem in text classification," *IEEE ,Third International Symposium on Information Processing,* 2010.

[5] M.V.Joshi, "Learning classifier models for predicting rare phenomena, [Ph.D.Thesis]." *University of Minnesota,* 2002.

[6] G.Weiss, " Explorations Special Issue on Learning from Imbalanced Datasets," *SIGKDD,* vol. 6, pp. 7-19, 2004.

[7] "A Comparative study on Feature Selection."

[8] G.E.A.P.A.Batista, R.C.Prati, and M.C.Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets* ,vol. 6, 2004.

[9] Y. Sun and e. S.Mohamed Kamel, "Cost-sensitive boosting for classification of imbalanced data. Pattern Rfecognition," vol. 40, pp. 3358-3378, 2007.

[10] E. Chen, Y. Lin, H. Xiong, Q. Luo, and H. Ma, "Exploiting probabilistic topic models to improve text categorization under class imbalance," *Information Processing & Management,* vol. 47, pp. 202-214, 2011.

[11] E. L. Iglesias, A. Seara Vieira, and L. Borrajo, "An HMM-based over-sampling technique to improve text classification," *Expert Systems with Applications,* vol. 40, pp. 7184-7192, 2013.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* pp. 321-357, 2002.

[13] S. Barua, M .M. Islam, X. Yao, and K. Murase, "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 26, pp. 405-425, 2014.

[14] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decision Support Systems,* vol. 48, pp. 191-201, 12// 2009.

[15] T. Lei and L. Huan, "Bias analysis in text classification for highly skewed data," in *Data Mining, Fifth IEEE International Conference on*, 2005, p. 4 pp.

[16] R. Navigli, "Word sense disambiguation:A Survey," *ACM Computing Surveys,* vol. 41, pp. 1-69, 2009.

[17] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems,* vol. 36, pp. 226-235, 2012.

[18] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," *Journal of Knowledge-Based Systems, Elsevier,* pp. 382-387, 2009.

[19] H. Amiri, A. AleAhmad, F. Oroumchian, C. Lucas, and M. Rahgozar, "Using OWA Fuzzy Operator to Merge Retrieval System Results," in *The Second Workshop on Computational Approaches to Arabic Script-based Languages*, USA,Stanford University, 2007.

[20] C. Saedi, Y. Motazadi, and M. Shamsfard, "Automatic translation between English and Persian texts," in *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages, Ottawa, Ontario, Canada*, 2009.

[21] T. Y. Liu, "Research on Feature Selection for Imbalanced Problem from Fault Diagnosis on Gear," *Advanced Materials Research,* vol. 466-467, pp. 886-890, 2012.

[22] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison-Wesley Longman Publishing Co., Inc., 1999.

[23] H. Guo and H .L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation," *The DataBoost IM Approach," ACM SIGKDD Explorations Newsletter,* vol. 6, pp. 30-39, 2004.

[24] Y. Sun, M. S. Kamel, a. A.K.C. Wong, and Y. Wang, " Cost-Sensitive Boosting for Classification of Imbalanced Data," *Pattern Recognition Letters,* vol. 40, pp. 3358-3378, 2007.

[25] H. He, Member, and E. A. Garcia, "Learning from Imbalanced Data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 21, 2009.

[26] M. Mike Wasikowski, IEEE and S. M. Xue-wen Chen, IEEE, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 22, 2010.