A Multi-Stage Framework of Textual Criminal Cases Categorization: Criminal and Legal Elements Approach

Sotarat Thammaboosadee
Technology of Information System Management Division,
Faculty of Engineering,
Mahidol University Thailand
sotarat.tha@mahidol.ac.th

ABSTRACT: *This paper proposes an identification framework of the possible criminal offences charges based on textual criminal cases of the Civil Law system. The framework is constructed as the model, devised as a multi-stage based on the defined charges structure in criminal law codes. The first stage is to modularly identify type of action which is designed based on the offences charges abstractly categorized by defined criminal elements. The second stage is to identify the additional legal elements, leading to general provisions which may affect to the sentence or amount of punishments. This classification stage is designed as multiple autonomous classification system. The integrated model is expected to be able to categorize charge type and provisional legal elements and to predict the final possible sentence and range of punishment. An evaluation aims to achieve high accuracy of classification while reserving explainable results, which is required in an application of legal domain.*

Keywords: Data Mining, Text Mining, Criminal Law, Criminal Elements, Legal Elements, Cases Categorization

Received: 17 July 2014, Revised 28 August 2014, Accepted 4 September 2014

1. Introduction

Most of the research of criminal law application of cases outcome prediction focuses on the Common Law system [1, 2, 3]. Hence, developing an intelligent agent, e.g. artificial intelligence, machine learning, and data mining, for the domain of the Civil Law system should be beneficial to users who interested in the legal system in various countries, including Thailand. The intelligent system of our interests in the Civil Law system will provide knowledge in terms of the validated textual case features to the possible offences charges and related legal elements to provide the final sentences which can lead to the theoretical possible range of punishments.

Hypothetically, in the judicial process of the Civil Law system, the facts of the case would be collected, investigated and induced into a set of related elements in law codes. According to this idea, there was a research article proposes a framework of two-stage data analysis [4] which are criminal facts, case diagnostic issues, and legal elements of the correspondingly applicable sections for identifying a range of possible sentences, ranges of a punishment and also their exacerbate or mitigation due to the additional provisions. The experimental results of this framework, scoped to the title 10 (offences against life and body) of Thai criminal codes, were successfully strengthened in both technical and legal aspects by its extensions [5, 6]. Technically, the framework was confirmed that the multi-stage identification system; multi-stage classifiers and also modularity work well in this application domain. Anyway, the case facts of this research series were initially collected in the well defined structure likes XML

(Extensible Markup Language) [7]. Therefore, some hidden valuable features or information in the textual case description may be ignored. Hence, this gap motivates author to apply the text analytics method such as text mining and natural language processing in the criminal charges categorization application.

Generally, the process of decision making in judicial process, specifically to the court level, starts from consideration in case facts and maps them into the criminal elements described in each offences charge. Afterwards, the additional legal elements identified in the general provisions book will be examined. This principle of consideration steps is a basis for this paper, supported by the designing of hierarchy and modularity classifier structure.

The remainder of the paper is organized as follows: the next section shows relevant past works in the data mining research in criminal law domain, especially in the text mining application in legal domain, and non-text mining application in cases prediction. Section three introduces the methodology in the classifying process which includes the overall architectural design. Finally, we conclude the paper and suggest the future research.

2. Related Works

According to the text analytics application in the legal domain, Ashley and Brüninghaus [8] proposed an automatic case text classification and outcomes prediction system. Their work focuses on the legal claim documents which are labeled by pre-defined 27 binary factors (positive and negative). Initially, a set of text processing methods is applied to the transform the raw text into the representative structure. Then several prediction algorithms are experimented and compared the accuracy of the classification model. Finally, the classifying results are indexed to retrieve the most similar case which is based for the outcome prediction. This work is quite similar to the proposed method of this paper. Anyway, their scope is the legal claims document which contains small size of factors. Thus, the flattened prediction models, with none of hierarchical or modularity, work well on classification task. Moreover, the aim of their proposed system is to retrieve the relevant and most similar document and ignore any explanations of the result. This is different from the requirement of an application in legal domain even if in text analysis system likes text mining. The reasoning or steps of factor determination should be retained for further argumentations.

Another similar related work was proposed by Francesconi and Passerini [9]. They proposed the provision classification in Italian legislative texts. Apart from the text preprocessing procedure, their experimental results show that the Multiclass Support Vector Machine (MSVM) [10] performs the best performance in term of classification accuracy for an overall pre-defined 11 classes. However, this work is quite strict to the provision which would be more complete if the offences are included. The integrated system of provision and offences classification system are also challenge in designing issues and legal interpretation.

Concerning to more technical in text mining method, Chen, Liu, and Ho [11] proposed the legal documents retrieval system, specific to criminal codes, for general public assisting purpose. Their special is that a query, consist of some common terms, input by non-lawyer users will be transformed and weighting mapped to legal terms assisting by commercial search engine. They scoped the input to 10 queries and proved their transforming method with information retrieval measurement. The advantage is that the cross-domain keywords transformation would be required in cases text categorization.

Apart from the text-based method, assume that the input data is already in well structuralizing. Yang et al. [12] constructed four independent classification models: random forest, decision tree, artificial neural network (ANN) [13] and support vector machine (SVM) [14] in a prediction system of an offender affiliation and murder victim. Their experiments showed that the SVM and ANN had the best accuracy in the training set. Unfortunately, they are worse in testing set because of overfitting problems.

Approaching to the more complete system, Stranieri and Zeleznikow [15] proposed the determination system of the final judgment for cases of Australia's family law domain, specific to the percentage of assets splitting for divorces. The ANN was applied to hierarchically discover the value of the defined legal factors, which was used to imitate the factors that the courts usually have to sequentially determine. The complete system was put online and interacted with users with the help of an induction rules method. However, this work is different from the domain of the work of this paper. The required final results in the Civil Law, especially in criminal codes, should specify all important legal elements based on the legal body and map onto charges.

In the next section, we will discuss the methodology for this system.

3. Methodology and Scope of Application

In this paper, a framework of multi-stage textual cases categorization model is proposed for the benefits of specifying prospective criminal elements and additional legal elements in criminal law, based only on the collection of verdicts of precedence in a specific sovereign Civil Law state, e.g. Thailand.

In this section, the author would like to describe the existing structure of the criminal codes used in this paper which is based for the modularity design. Then the designing of text preprocessing procedures and both classification systems, which are offence charges and general provisions identifier, are demonstrated respectively. Finally, the method of integration is proposed.

3.1 Criminal Codes Structure and Consideration Process

According to the criminal codes of Thailand [16], which are in the same fashion in other Civil Law system country, overall law codes are separated into two parts (or "*books*" in legal term). The first book, general provisions, describes the definitions, theories, principles, and deposition rules applied in all offences. The general provisions are separated into two subparts (or "titles") which consist of nine subparts (or "*chapters*") and cover the law codes (or "*sections*") no. 1-106. The second book is specific offences. This book specifies the criminal offences in term of actions and responsible punishments. Based on the similarity of crimes the specific offences charges are categorized into 13 titles, and divided into chapters in some titles, which cover section no. 107-398. The list of books, titles, and chapters is shown in Table 1.

Anyway, some chapters are out of scopes of the proposed framework. In Title 1, Chapter no. 1-3 is excluded since they describe definitions and principles which rarely involve in charges identification procedure. Moreover, since the proposed methodology is based on data mining techniques which are needed for availability of the data [17]. There are some Titles and Chapters in the second Book are eliminated because of its rare occurrence. The ignored sections are: Title 1, Chapter 2 of Title 3, Title 4, Title 5, Chapter 1 of Title 7, Chapter 2 of Title 7, Title 10 Chapter 3, Title 10 Chapter 4 and Chapter 2 of Title 11 since they are less than 0.5 % of occurrence of approximated 19,000 total cases in preliminary study.

According to the process of case consideration, the court usually firstly determines the offence charges by identifying a set of criminal elements specified in each offences section. If one or more offences charges are satisfied, the additional legal elements in general provision books will be then determined and may change the sentence or punishment. This determining step is based for the proposed framework which the offences should be specified before the general provision, as shown in Figure 1.

Anyway, some chapters are out of scopes of the proposed framework. In Title 1, Chapter no. 1-3 is excluded since they describe definitions and principles which rarely involve in charges identification procedure. Moreover, since the proposed methodology is based on data mining techniques which are needed for availability of the data [17]. There are some Titles and Chapters in the second Book are eliminated because of its rare occurrence. The ignored sections are: Title 1, Chapter 2 of Title 3, Title 4, Title 5, Chapter 1 of Title 7, Chapter 2 of Title 7, Title 10 Chapter 3, Title 10 Chapter 4 and Chapter 2 of Title 11 since they are less than 0.5 % of occurrence of approximated 19,000 total cases in preliminary study.

According to the process of case consideration, the court usually firstly determines the offence charges by identifying a set of criminal elements specified in each offences section. If one or more offences charges are satisfied, the additional legal elements in general provision books will be then determined and may change the sentence or punishment. This determining step is based for the proposed framework which the offences should be specified before the general provision, as shown in Figure 1.

3.2 Textual cases preprocessing

Similar many Asian languages, the Thai text in case verdicts have to be first segmented and applied as features. Specifically to legal domain, the text segmentation procedure should be dictionary-based method [18] since the documents consist of several legal terms. An over-segmented lexicon may cause of losing in features in legal terms. Additionally, each separated term may have a different linguistic role which leads to different implication. The part-of-speech tagger [19] should be labeled to the segmented terms. Therefore, same word may be threatened as two different features. Each separated feature should be transformed to numerical format. Among the candidates the term frequency-inverse document frequency (TD-IDF) [20] should be the most appropriate one since it concerns both inter and intra occurrence of words.

3.3 Offences Charges Classification

Based on the numerical transformed features generated by described text preprocessor, the offences charges have to be

Book 1. General Provisions	
Title 1. Provisions Applicable to General Offences	Chapter 6. Principals and Supporters
Chapter 1. Definitions	Chapter 7. Concurrence of Offences
Chapter 2. Application of Penal Laws	Chapter 8. Recidive
Chapter 3. Punishments and Measures of Safety	Chapter 9. Prescription
Chapter 4. Criminal Liability	
Chapter 5. Attempt	
Book 2. Specific Offences	
Title 1. Offences Relating to the Security of the Kingdom	
Chapter 1. Against the King, the Queen, the Heir-apparent and the Regent	Chapter 3. Against the External Security of the Kingdom
Chapter 2. Against the Internal Security of the Kingdom	Chapter 4. Against the Friendly Relations with Foreign States
Title 1/1. The Offence in Respect of Terrorization	
Title 2. Offences Relating to Public Administration	
Chapter 1. Against Officials	Chapter 2. Malfeasance in Office
Title 3. Offences relating to the Justice	
Chapter 1. Against the Judicial Officials	Chapter 2. Malfeasance in Judicial Office
Title 4. Offences relating to Religion	**Title 5. Offences relating to Public Peace**
Title 6. Offences relating to Causing Public Dangers	**Title 7. Offences relating to Counterfeit and Alteration**
Chapter 1. Relating to Currencies	Chapter 3. Relating to Document
Chapter 2. Relating to Seals, Stamps and Tickets	Chapter 4. Relating the Electronic card
Title 8. Offences Relating to Trade	**Title 9. Offences Relating to Sexuality**
Title 10. Offences against Life and Body	
Chapter 1. Offences Causing Death	Chapter 3. Abortion
Chapter 2. Against Body	Chapter 4. Abandonment of Children, Sick Persons or Aged Persons
Title 11. Offences against Liberty and Reputation	
Chapter 1. Against Liberty	Chapter 3. Defamation
Chapter 2. Disclosure of Private Secrets	
Title 12. Offences against Property	
Chapter 1. Theft and Snatching	Chapter 5. Misappropriation
Chapter 2. Extortion, Blackmail, Robbery and Gang-Robbery	Chapter 6. Receiving Stolen Property
Chapter 3. Cheating and Fraud	Chapter 7. Mischief
Chapter 4. Cheating Against Creditors	Chapter 8. Trespass
Title 13. Petty Offences	

Table 1. Category of all Sections in Thai Criminal Codes

identified first. Anyway, each criminal case probably able to be matched with more than one offence charge. For example, an offender who committed a trespassing crime might also commit a homicide. Thus, this classification stage should not be performed as a single classification system and lead us to consider a one-class classification system. However, the one-class

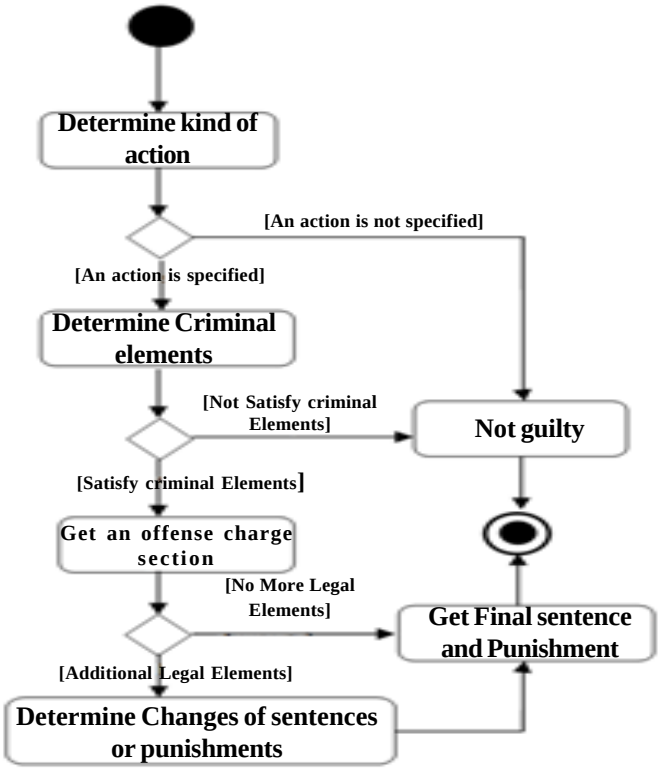

Figure 1. Charges Determination Process

classifier for all offence charges should cause an over fitting for each crime which is typically nearby especially the offences in the same Chapter. Moreover, some sections contain the same crime but different additional legal elements, e.g. section 288 refers to the homicide with intention and section 290 refers to the same crime but with the negligence element instead. Accordingly, the identification system should be applied just in Title or Chapter level, based on criminal law codes, and should be in hierarchical structure. The Title of crimes has to be former identified and then lead to the Chapter categorizing. According to these concepts and scope of application as described in section 3A, an architectural design of offences charges classification system is shown in Figure 2.

Approaching to the applied algorithm in this classifier stage, although there are several researches suggested the Support Vector Machine (SVM) and Artificial Neural Network (ANN) should be appropriate for text mining application.

Theoretically, the SVM classification algorithm [13], is to find a separation between hyper planes defined by classes of data. Its goal is to discover the largest margin of separation of the data. Therefore, the SVM algorithm can avoid falling into the trap of local optimality and operate well even in moderately large feature (or attributes) sets. On the other hands, the ANN is a computational model that is inspired by the structural and functional aspects of the biological neural network system. It consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach. There are several types of ANN. In this research, the ANN with feed-forward and back-propagation architecture [14] is chosen. Anyway, since the traditional SVM and ANN works as a black-box, their results typically lack of interpretation. To eliminate this limitation, the authors apply a modular architecture to both classification methods. The modularity is designed into a multi-classifier of specialized identifiers. Once a result is obtained from the system, one can trace back the echelon to derive the interpretation. According to the set up of an experiment in this research, both classification algorithms should be performed and compared the categorizing performance.

3.4 General Provisions Charges Classification

As stated, the scoped general provision charges are in Chapter 4-9. Initially, there is previous research proposed the data structure for criminal law codes collection, represented in ontology [21]. Based on the proposed ontology, the classes of selected legal elements according to selected general provision are listed in Table 2. Noted that only the provisions those involve in charges determination are selected.

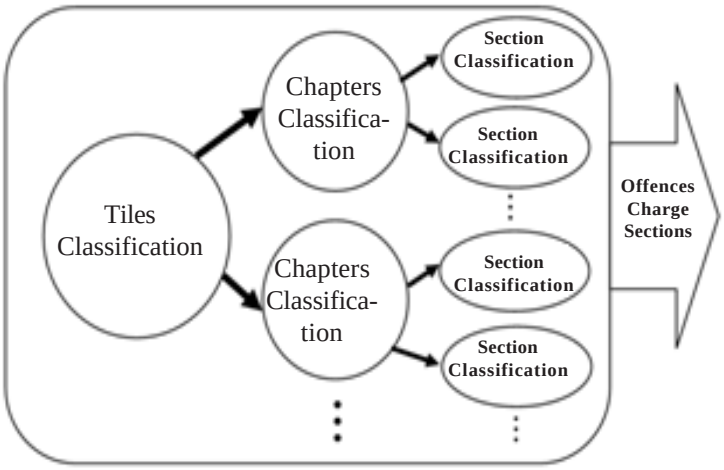

Figure 2. An Architectural Design of Offences Charges Classification System

Class name	Class values
Age	Under_10, Under_15, Under_18, Under_20, Adult
Anger	Anger, No_Anger
Awareness	No_awareness, Moderately_aware, Fully_aware
Commitment	Prepare, Uncommitted_by_employer, Intervention, Attempt, Attemp_and_desist, Unachievable_attempt, Success
Intention	Intention, Negligence, No_intention
Instigator	Instigation, No_instigation
Prevention	No_prevention, Prevent_from_harm, Necessary_prevention, Over_excess_prevention

Table 2. List of Selected Legal Elements according to General Provision Charges

Technically, each legal element is separately considered and this leads the design of independent classification model. The one-class classifiers like SVM and ANN are still good candidates in experiments.

3.5 An Elements-based Multi-Stage Charges Identification Model

According to the proposed methodologies, their integration, an elements-based architecture of multi-stage charges identification model is shown in Figure 3.

Firstly, the textual case verdicts are processed and transformed into features based on the methodology proposed in section 3B. Then, the offences charges are identified hierarchically and modularly based on the criminal codes structure and described in section 3C. Consequently, the additional legal elements based on general provisions sections are independently classified as described in section 3D. In some chapters of offence charges, the additional legal elements may change the offence charges from the initial one to be its extension. For example, a homicide crime is map as section 288 may be considered as section 297 instead if the court can prove that the offender performed an action with negligence, not intention. Hence, the resulting offence charges have to be mapped with legal elements to determine the final charges list.

Finally, to further squeeze the performance, the experiment is planned to be executed under the 10-fold cross validation method and data will be cleansed for ensuring integrity [22]. Furthermore, the hold-out validation method is also considered to guarantee the usage in real application.

5. Conclusion

A framework of elements-based multi-stage textual charges classification model is proposed to identify criminal charges according

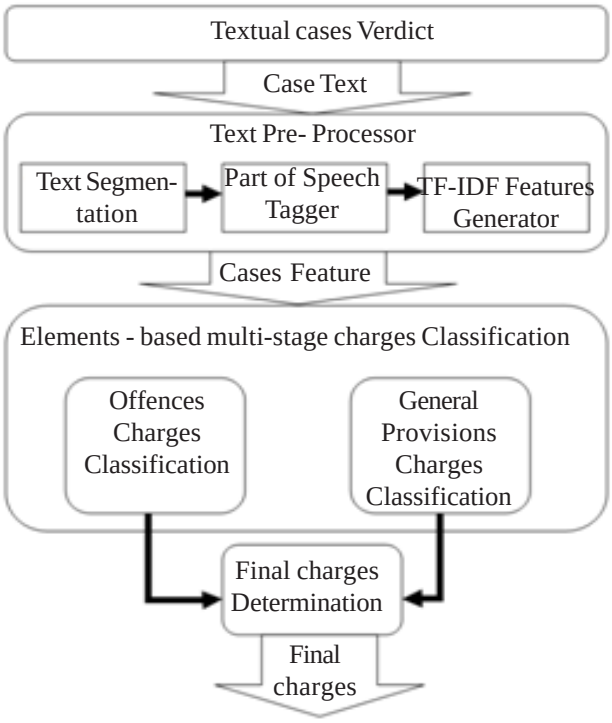

Figure 3. An Elements-based Multi-Stage Charges Identification Model

to the legal domain of interest which is the criminal law in the Civil Law system which the sentences are relied on law sections. To support the design of the model, the classifiers are separated into 2 levels based on the real charges determination process: offence charges identification and general provision charges identification. The first classification stage hierarchically identifies charges based on the defined criminal elements in the criminal codes. The second stage, general provision charges (or legal elements) identification, is designed as independent model and are used to mapped with the offences charges in order to identify the final charges. Since the input data is the textual verdict, the required text-preprocessor methods include text segmentation and part-of-speech tagger. Both classification stages are applied with the Support Vector Machine (SVM) and Artificial Neural Network (ANN) with one-class classifier architecture. The framework also considers the high dimensionality by text preprocessors handling standardization, validation, and most importantly the factor analysis to dynamically and context-sensitively decrease the input dimensions.

At the current stage, the framework is in the process of implementing into a complete categorization system. Once finished, the performance of the system will be measured in terms of model accuracy and contingency analysis.

References

[1] Apple, J. G. Deyling, R. P. (1995). A Primer on the Civil-Law System, Federal Judicial Center, Washington D.C.

[2] Bilgi, N. B., Kulkarni, R. V. (2008). An Investigative Survey of Application of Knowledge Based System in Legal Domain, *Int. J. of Information Technology and Knowledge Management*, p. 517-525.

[3] Ashley, K. D. (2004). Case-Based Models of Legal Reasoning in a Civil Law Context, Int. Congress of Comparative Cultures and Legal Systems of the Instituto de Investigaciones Juridicas, Universidad Nacional Autonoma de Mexico, Mexico City.

[4] Thammaboosadee, S., Watanapa, B., Charoenkitkarn, N. (2012). A Framework of Multi-Stage Classifier for Identifying Criminal Law Sentences, Procedia Computer Science, 3, p. 53-59.

[5] Thammaboosadee, S., Watanapa, B. (2013). Identification of Criminal Case Diagnostic Issues: a Modular ANN Approach, *Int. J. of Information Technology and Decision Making*, 12 (3).

[6] Thammaboosadee, S., Watanapa, B., Chan, H., Silparcha, U. (2014). A Two-Stage Classifier That Identifies Charge and Punishment under Criminal Law of Civil Law System, *IEICE Transactions*, 97-D (4), p. 864-875.

[7] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. (2010). Extensible markup language (XML) 1.0, W3C Recommendation, http://www.w3.org/TR/REC-xml, accessed October.

[8] Ashley, K. D., Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law*, 17 (2) p. 125-165.

[9] Francesconi, E., Passerini, A. (2007). Automatic Classification of Provisions in Legislative Texts, I*nt. J. on Artificial Intelligence and Law*, 15 (1) 1–17.

[10] Duan, K. B., Keerthi, S. S. (2005). Which Is the Best Multiclass SVM Method? An Empirical Study, Lecture Notes in Computer Science, 3541, p. 278.

[11] Chen, Y. , Liu, Y., Ho, W. (2013). A text mining approach to assist the general public in the retrieval of legal documents, J. of the Association for Information Science and Technology, 64 (2), p. 280-290.

[12] Yang, R., Olafsson, S. (2011). Classification for Predicting Offender Affiliation with Murder Victims", Expert Systems with Applications, 38 (11), p. 13518-13526.

[13] Craven, M., Shavlik, J. (1997). Using Neural Networks for Data Mining, Future Generation Computer Systems, 13, p. 211-229.

[14] Cortes, C., Vapnik, V. (1995). Support-vector networks, Machine Learning, 20 (3), p. 273, 1995.

[15] Stranieri, A., Zeleznikow, J. (2005). Knowledge Discovery from Legal Databases, Law and Philosophy Library, 69, Springer.

[16] Yuthankun, Y. (2009). The criminal code: translated Thai-English, Soutpaisal Publisher, Bangkok.

[17] Witten, I. H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann.

[18] Haruechaiyasak, C., Kongthon, A. (2013). LexToPlus: A Thai Lexeme Tokenization and Normalization Tool, Int. Joint Conf. on Natural Language Processing, p. 9–16.

[19] Charniak, E. (1997). Statistical Techniques for Natural Language Parsing, AI Magazine, 18 (4), p.33–44.

[20] Salton, G., McGill, M. J. (1986). Introduction to modern information retrieval, McGraw-Hill.

[21] Thammaboosadee, S., Watanapa, B. (2013). Criminal Law Ontology for Identifying Possible Sentences from Specific Legal Elements, Law & Practice: Critical Analysis and Legal Reasoning, p. 398-408.

[22] Devijver, P. A., Kittler, J. (1982). Pattern Recognition: A Statistical Approach, Prentice-Hall, London.