UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

ESTUDIO DE MODELO DE TÓPICOS APLICADO A TRANSCRIPCIONES DE CLASES DE MATEMÁTICAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

ROBERTO EDUARDO CASTILLO NAVARRO

PROFESOR GUÍA:
PABLO DARTNELL ROY

MIEMBROS DE LA COMISIÓN:
RAFAEL CORREA FONTECILLA
ROBERTO ARAYA SCHULZ

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: ROBERTO EDUARDO CASTILLO NAVARRO
FECHA: 2016
PROF. GUÍA: SR. PABLO DARTNELL ROY

ESTUDIO DE MODELO DE TÓPICOS APLICADO A TRANSCRIPCIONES DE CLASES DE MATEMÁTICAS

El presente trabajo surge de la necesidad de explorar y luego evaluar métodos de clasificación para transcripciones de clases de matemáticas, en el contexto de la investigación del quehacer docente desde un punto de vista cualitativo y cuantitativo realizada desde el Centro de Investigación Avanzada en Eduación (CIAE). En su primera sección se encuentra una breve descripción de las técnicas utilizadas y a explorar. El segundo capítulo enumera y describe posibles metodologías para evaluar el desempeño de las técnicas de clasificación a utilizar u otras que podrían implementarse a futuro. Se observan algunas de sus características, con el fin de entender sus méritos como herramientas de evaluación y justificar la elección de alguna de ellas. Finalmente se presentan los resultados obtenidos al aplicar la metodología de evaluación a datos reales correspondientes a transcripciones de grabaciones de clases proporcionadas en forma voluntaria por docentes que decidieron colaborar con la investigación realizada por CIAE.

A Isabel, porque me enseñas a redescubrir el mundo cada día.

Agradecimientos

En primer lugar, quisiera agradecer a mi profesor guía el señor Pablo Dartnell, por su apoyo, confianza, paciencia y consejo en el desarrollo de este trabajo de memoria de título y también en su rol como académico del departamento de ingeniería matemática, pues es para mí un ejemplo y un mentor tanto en lo profesional como en lo personal.

También quisiera extender mi agradecimiento al resto del cuerpo académico del departamento de ingeniería matemática, por la formación de excelencia que recibí de ellos, sin que el profesionalismo dejara de lado el respeto y la calidad humana.

Además agradezco a los funcionarios del departamento y de la facultad de ciencias físicas y matemáticas, pues su labor silenciosa posibilita diariamente el ambiente fecundo de cultivo intelectual que tanto valoramos.

Adicionalmente agradezco a mis compañeros estudiantes del Departamento de Ingeniería Matemática, por el crecimiento conjunto, la sana competencia, las discusiones fructíferas en torno a problemas comunes e individuales y, por supuesto, las instancias de sano esparcimiento que compartimos.

Un agradecimiento especial a mi familia, por la paciencia, consejo, apoyo y aliento en cada momento y en cada dificultad y tropiezo. El apoyo y comprensión por los descuidos y malabares asociados a intentar ser muchas cosas a la vez es invaluable y solo me queda intentar ser tan paciente y comprensivo con cada uno de ellos como lo han sido y seguramente seguirán siendo conmigo.

Finalmente quisiera agradecer el financiamiento del Proyecto Fondef D11I1009 y del programa PIA/Basal de Conicyt a través del proyecto FB0003, pues este apoyo es de vital importancia para el desarrollo de la investigación dentro de la cual se enmarca y desarrolla el presente trabajo de memoria.

Tabla de Contenido

Índice de Ilustraciones

Introducción

Una meta fundamental de alcanzar para el desarrollo futuro de un país es el logro de aprendizajes significativos en matemática y ciencia, por lo que constantemente se busca métodos que permitan fortalecer y enriquecer la labor docente. Continuando con este trabajo de investigación de las interacciones entre estudiantes y docentes, de la que se tiene registro desde inicios del siglo XX, el Centro de Investigación Avanzada en Educación (CIAE) prepara diversos sistemas, mecanismos y mediciones que permitan generar progreso en esa dirección. Uno de los proyectos del CIAE es la creación de un sistema de auto soporte para el mejoramiento de las prácticas docentes a través del proyecto FONDEF D11I1009 y dentro de este marco es que se realiza el presente trabajo de título.
Antecedentes para este proyecto comienzan desde la observación simple del conteo de palabras o interacciones entre estudiantes y el docente [10] y a través del tiempo ha avanzado a recolectar información sobre prácticas pedagógicas específicas en el aula [9], los resultados de estos esfuerzos produjeron avances ciertamente, pero sigue la inquietud por lograr mejoramientos sustanciales más allá de casos específicos, lo que constituye un desafío mayor para lograr la optimización de la eficiencia, efectividad y la atención de los estudiantes [7]. Con el fin de calificar apropiadamente una clase de manera automática, existen diversos factores que han de ser tomados en cuenta, posiblemente una cantidad excesiva de variables y contextos demasiado diversos como para abordar el problema con generalidad. Sin embargo, entregar información oportuna respecto de indicadores específicos presentes en una clase puede ser de gran utilidad para un docente que busca retroalimentación para lograr enriquecer su práctica de acuerdo a metas y objetivos según el contexto particular en que se desenvuelve. Es así que proveer al profesor de retroalimentación puede fortalecer y enriquecer la práctica de la enseñanza. Existen diversos métodos para entregar retroalimentación al profesor, tales como la observación en aula o la revisión de videos de sus clases. En particular, actualmente las prácticas docentes son generalmente monitoreadas por observadores entrenados que utilizan rúbricas para identificar y clasificar situaciones que se dan en las salas de clases, o revisar videos previamente grabados de las clases [8]. Sin embargo, este proceso es lento, tedioso y puede presentar errores, haciendo que sea prácticamente imposible de utilizar para entregar de forma masiva y permanente a los docentes sugerencias inmediatas de estrategias de enseñanza en el asesoramiento del desempeño y necesidades de aprendizaje de los estudiantes [1].
Teniendo en cuenta lo anterior, el proyecto de CIAE ha buscado la producción de una herramienta generadora de información y retroalimentación constante e inmediata, con el fin de apoyar la labor docente a través de la autoobservación. Con este objetivo en mente, dado que se busca generar un sistema accesible, se ha propuesto un sistema que comience por la captura del audio del profesor en la clase a través de una aplicación para *Smartphone*, lo que permite contar con el contenido verbal de la clase para su estudio y análisis posterior.

El sistema funciona de la siguiente forma: 1) La aplicación registra la grabación del audio de una clase, 2) Dicha grabación se transcribe automáticamente a texto, y 3) Finalmente, el texto se analiza para identificar y cuantificar la presencia de los factores en observación. En esta oportunidad se considera la medición sobre seis contenidos matemáticos, presentes en el currículum nacional, y la de diecisiete prácticas docentes, escogidas por existir amplio acuerdo en su importancia como indicadores de calidad [1] [7].

En este punto encontramos dos temas que son el objeto de estudio de este trabajo de título. En primer lugar se encuentra la elección de un algoritmo de clasificación adecuado para identificar y señalar apropiadamente la presencia de los factores en observación que se han considerado dentro del proyecto, el capítulo 1 ha de presentar una discusión de algunas alternativas que se presentan como una solución para esta parte del problema. En segundo lugar, una vez escogido el método de clasificación, se ha de escoger un método que permita evaluar su efectividad, el capítulo 2 presenta una revisión de algunos posibles indicadores que sirvan para cuantificar adecuadamente el desempeño de la clasificación. Finalmente, el capítulo 3 describirá los resultados conseguidos por el proyecto y a la luz de ello se ha de discutir las fortalezas y debilidades del sistema en su estado actual.

Capítulo 1

Clasificación De Textos

Con el objetivo de generar un mecanismo de información al docente, el primer paso es generar un sistema automatizado que permita identificar la presencia o ausencia de determinados contenidos o prácticas pedagógicas en sus clases.
Para esto es necesario analizar la transcripción del discurso verbal de la clase. Se asume en este trabajo que se cuenta con transcripciones de buena calidad de las grabaciones de las clases (pese a que la obtención de este material puede ser un desafío importante en sí mismo), a partir de este punto es necesario entonces discutir los méritos de los métodos considerados para la clasificación.
En la tarea que se presenta, los datos a considerar son las palabras presentes en el texto. Se observa la necesidad de eliminar posibles fuentes de ruido en las transcripciones, dado que los métodos de clasificación se encargan de identificar y utilizar las palabras o conceptos más utilizados para realizar inferencia es necesario depurar las entradas a fin de lograr mejores resultados. Es por esto que se realiza un proceso de limpieza de los datos que considera ignorar palabras que no tengan que ver con contenidos ni prácticas docentes, tales como artículos, pronombres o conectores y agrupar las palabras restantes en 'familias' según su raíz.
Posteriormente, para cada categoría de clasificación que se desee implementar, el algoritmo de clasificación debe considerar cuántas veces se repiten las diferentes raíces en cada texto por separado, y a partir de clasificaciones manuales que se usan como datos de entrenamiento, le asigna a cada raíz de cada texto un peso (importancia), generando así una bolsa de palabras que se asocian a cada categoría de clasificación.
En el proyecto actualmente en desarrollo se ha optado por usar el algoritmo de clasificación Random Forests [4].

1.1. Algoritmo Random Forests

El algoritmo *Random Forests* consiste en una combinación de árboles predictores generados a partir de vectores aleatorios independientes e idénticamente distribuidos. El método permite obtener clasificaciones con errores pequeños al utilizar la información obtenida de las clasificaciones independientes generadas por cada uno de los árboles.

Este tipo de métodos tiene como idea fundamental 'promediar' muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la variación. Los árboles son los candidatos ideales para este proceso, dado que ellos pueden registrar estructuras de interacción compleja en los datos, y si crecen suficientemente profundo, tienen relativamente baja parcialidad. Producto de que los árboles son notoriamente ruidosos, ellos se benefician grandemente al promediar.

Cada árbol es construido siguiendo un esquema como el que sigue:

- Sea N el número de casos de prueba, M es el número de variables en el clasificador.
- Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado; m debe ser mucho menor que M
- Elegir un conjunto de entrenamiento para este árbol y usar el resto de los casos de prueba para estimar el error.
- Para cada nodo del árbol, elegir aleatoriamente m variables en las cuales basar la decisión. Calcular la mejor partición a partir de las m variables del conjunto de entrenamiento.

Para la predicción un nuevo caso es empujado hacia abajo por el árbol. Luego se le asigna la etiqueta del nodo terminal donde termina. Este proceso es iterado por todos los árboles en el ensamblado, y la etiqueta que obtenga la mayor cantidad de incidencias es reportada como la predicción.

Este es el método que dentro del proyecto se decidió utilizar debido a su buen desempeño y buenas propiedades de convergencia y exactitud que ha exhibido Breiman en su publicación al respecto.
A pesar de que al revisar la literatura al respecto random forests exhibe buenas propiedades que lo perfilan como una excelente herramienta, este trabajo surge inicialmente con la motivación de estudiar una alternativa a este método de clasificación, el cual corresponde al modelo de tópicos.

1.2. Modelo de tópicos

Los modelos de tópicos corresponden a una familia de algoritmos diseñados para explorar y descubrir los temas subyacentes en una colección de documentos que *a priori* carece de una estructura clara, aplicados cuidadosamente permiten encontrar relaciones o grupos de palabras (tópicos) que permiten entender la naturaleza de una gran colección de documentos. En el caso de este proyecto la colección corresponde a las transcripciones de clases de matemática y el objetivo sería descubrir a través de los tópicos la existencia de las prácticas docentes buscadas u otros temas de interés para el diagnóstico y retroalimentación del desempeño de un profesor participante.
El modelo de tópicos más utilizado y que sería aplicable al caso presente es el de *Latent Dirichlet Allocation* (LDA), el cual es un modelo generativo probabilístico para colecciones de datos discretos, como un conjunto de textos. El LDA maneja el modelo bayesiano jerár-

quico, en el que cada elemento de un corpus de texto es modelado como una mezcla finita, sobre un conjunto fundamental de tópicos, donde a su vez cada tópico se modela como una mezcla infinita sobre un conjunto subyacente de tópicos probables, es decir, que todos los corpus se representan como mezclas aleatorias, sobre tópicos ocultos. En el modelo LDA cada documento es visto como una mezcla de varios tópicos, cuya distribución se supone de tipo Dirichlet. En la práctica, esto resulta en mezclas coherentes de tópicos en un documento.

Por ejemplo, un modelo LDA podría tener los tópicos GATO y PERRO. El tópico GATO tiene probabilidades de generar varias palabras: leche, maullido, gatito, por lógica la palabra gato tendrá la probabilidad más alta dado este tópico. Por otro lado, el tópico PERRO tiene la probabilidad de generar las palabras: cachorro, ladrido, hueso, y esta última podría tener una alta probabilidad. Las palabras sin determinada relevancia, tendrán aproximadamente la misma probabilidad entre sus clases (o pueden ser colocadas en una categoría aparte).

A pesar del potencial que este modelo de tópicos pareciera exhibir, intentos por utilizarlo en el problema actual no generaron buenos resultados a causa de la falta de dirección en la búsqueda de los temas unificadores de los tópicos, por lo que no resulta una alternativa a la buena clasificación lograda por *Random forests*.

Capítulo 2

Indicadores para la evaluación de calidad de los clasificadores

Una vez que se han considerado los métodos de clasificación, aparece la segunda parte del problema en estudio: determinar su utilidad y confiabilidad. Para evaluar adecuadamente el desempeño de los clasificadores en primer lugar es necesario decidir que tipo de mediciones hemos de considerar como indicaciones de buen rendimiento.
Teniendo en consideración el contexto del proyecto en que se enmarca este trabajo, es importante que el clasificador alcance un desempeño equivalente al de un experto humano que realizara la misma tarea, pues uno de los objetivos a largo plazo es el de complementar una retroalimentación humana por este sistema, lo que permite a un docente recibir información expedita y, por lo tanto, realizar ajustes en su labor de acuerdo al informe recibido desde el sistema.
Se han de escoger indicadores que permitan comparar dos clasificadores o evaluadores, por ejemplo: un experto humano y un clasificador automatizado o el grado de acuerdo entre dos expertos. En la literatura existen variados enfoques posibles para realizar esta tarea, y este capítulo se centra en la discusión de algunos de ellos, sus virtudes e inconvenientes además de fundamentar la elección de uno de ellos para el desarrollo del proyecto.

2.1. Acuerdo entre clasificadores

Al momento de cuantificar el acuerdo entre evaluadores o clasificadores es claro que se requiere desarrollar un criterio objetivo y robusto para determinar si existe acuerdo entre los observadores y, más aún, determinar el grado de acuerdo en los casos en que existe: ¿existe suficiente acuerdo como para sustituir un evaluador por el otro y obtener los mismos resultados?, ¿alguno de los evaluadores comete 'más errores' que los demás?, ¿alguno de los evaluadores es más o menos estricto?
Afortunadamente, un problema similar ha sido tratado anteriormente por la literatura biomédica, debido a que en esta área es frecuente que dos profesionales den diagnósticos disímiles usando la misma información disponible, por lo que es importante evaluar la confiabilidad

del diagnóstico para poder garantizar el tratamiento correcto a pacientes.
Existen distintos indicadores que permiten evaluar la robustez del acuerdo entre evaluadores o clasificadores, a continuación se presentan algunas de las posibilidades disponibles y se avanza a fundamentar la elección de la que se ha utilizado en este trabajo.

2.2. Indicadores de confiabilidad entre evaluadores

La confiabilidad entre clasificadores (*inter-rater reliability* es el concepto original en inglés) es un concepto que puede resultar particularmente adecuado para la tarea que se ha de llevar a cabo. Esta idea consiste en definir una razón o índice que cuantifica el grado de acuerdo entre dos clasificadores. En el trabajo presente, representa una medida del acierto que alcanza el sistema automatizado de clasificación al ser comparado con expertos humanos realizando la misma tarea.

2.2.1. Comparación del número de aciertos

Un primer indicador es una métrica simple que consiste en contar el número de coincidencias entre las evaluaciones. Pese a que aparentemente este indicador podría ser de utilidad, pronto se hace evidente que en realidad es una mirada bastante incompleta del grado de acuerdo o concordancia entre evaluadores, pues quizás sería preferible que en clasificaciones ordinales (evaluaciones o similares) los evaluadores siempre tengan opiniones cercanas, aunque no sean exactamente coincidentes, a la situación en que hay acierto en varios casos, pero en otros las diferencias son abismales.

Es por eso que se busca otra alternativa que refleje mejor la similitud en los criterios de elección, un concepto bastante útil que aparece es medir una especie de probabilidad de acuerdo, teniendo en cuenta que la clasificación de cada uno de los evaluadores se puede considerar una variable aleatoria, existe el desarrollo de coeficientes que permiten estimar la confiabilidad o 'intercambiabilidad' de dos o más evaluadores de acuerdo a las coincidencias que éstos tienen.

2.2.2. Coeficientes con corrección de acuerdo por azar

Dado lo anterior y con el objetivo de buscar un indicador que de cuenta del verdadero grado de acuerdo entre clasificadores, es importante tener en mente la solución a problemas similares, situaciones en que se busca comparar el desempeño o el criterio de dos o más clasificadores o evaluadores ante una misma decisión con la misma evidencia.
Es natural que en ocasiones ocurran coincidencias de decisión por azar más que por una similitud o acuerdo inherente a la metodología de ambos evaluadores y este tipo de acuerdo accidental debería ser eliminado de las consideraciones de acuerdo. Al buscar profundizar

en esta idea se encuentran varios coeficientes cuyo espíritu es considerar el porcentaje de acuerdos respecto del total, un estimador de la probabilidad de acuerdo, pero introduciendo un término de corrección, el cual corresponde a una estimación de la probabilidad de que el acuerdo que se produce se deba simplemente a una coincidencia accidental.
Existe una familia de indicadores o coeficientes de acuerdo corregidos por azar que vienen dados por la fórmula

$$\kappa = \frac{p_a - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}}$$

Donde p_a representa la probabilidad de acuerdo, y p_{e} representa la probabilidad de acuerdo obtenido accidentalmente. Dependiendo de la forma en que se estiman estas magnitudes, el coeficiente tiene distitas designaciones. Algunos de los coeficientes en esta familia son Kappa de Cohen, Alpha de Krippendorff, Pi de Scott o el AC1 de Gwet, este último es la elección a utilizar en este trabajo para la evaluación del desempeño de los clasificadores, debido a que es el que tiene comportamientos más robustos ante muestras desequilibradas en la distribuciones de clasificaciones, que es el caso que se encuentra en los datos de las transcripciones de clases, donde hay categorías en que los casos en que se observa la práctica son notablemente minoritarios. Una exhibición detallada acerca del mejor desempeño de AC1 en estos casos en comparación a otros coeficientes puede encontrarse en *Handbook of inter-rater Reliability* de Gwet [5].

En el caso particular de AC1, se tiene que $p_{\mathrm{e}} = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k(1 - \pi_k)$ donde π_k representa la probabilidad de que un evaluador clasifique a un sujeto en la categoría k y q representa el número total de categorías disponibles. En el contexto de este trabajo, para cada uno de los contenidos o prácticas docentes que se busca identificar hay dos categorías posibles ($q = 2$): el contenido o práctica se observa en el texto, o bien no se observa evidencia en el texto de su presencia.

Capítulo 3

Resultados en el contexto del proyecto

3.1. Datos Utilizados

Para crear y entrenar el modelo de transcripción automática y el modelo de clasificación de la aplicación, 93 profesores registraron el audio de un total de 866 clases para este proyecto. Todos los registros de audio de clases fueron grabados, transcritos y clasificados a través de la aplicación móvil del proyecto. De cada audio de clase se utilizó un extracto de cinco minutos de duración, escogido de manera tal que fueran los cinco minutos con más palabras habladas de toda la clase (a través de una aplicación móvil, de forma automatizada). Se ha utilizado audio registrado usando micrófonos alámbricos de manos libres conectados al celular.
Los audios de cinco minutos fueron transcritos manualmente y luego expertos identificaron, en cada uno de dichos segmentos de clase, la presencia o ausencia de cada práctica y contenido buscado, apoyándose tanto en los textos como los audios para las clasificaciones. De los 866 registros de audio utilizados para entrenar el sistema, 309 clases fueron transcritas y clasificadas por tres expertos contratados para ello, y luego las siguientes 557 fueron transcritas y clasificadas por docentes que estaban utilizando la aplicación.
Para las transcripciones, los transcriptores escuchaban el audio, y debían ir corrigiendo la transcripción automática preliminar propuesta por una versión inicial del sistema, dejándola exactamente igual al audio de la clase.
Los contenidos matemáticos eran: álgebra, aritmética, ecuaciones, fracciones, geometría y proporcionalidad; y las prácticas eran (observar que son diferentes, pero no necesariamente mutuamente excluyentes entre ellas): aclaración de conceptos, realización de cálculos, clasificaciones de conceptos o contenidos, crear buen ambiente en el aula, dinamismo en la exposición, generar espacio para la discusión, trabajo con ejemplos, presentación de explicaciones, hacerle preguntas a los estudiantes, dar instrucciones, interacción con estudiantes, interpelación a estudiantes, nombrar a los estudiantes por sus nombres, desarrollo de razonamiento matemático, reforzamiento de contenidos o habilidades previamente cubiertas, y valoraciones negativas y positivas de los aportes o trabajo de los estudiantes. Para cada categoría, tanto de contenidos como de prácticas docentes, usando una plataforma web especialmente construida para estos efectos, los clasificadores deben identificar, para cada texto a clasificar, si la categoría está presente o no.

Para la transcripción de los audios a texto, se contó con el apoyo de colaboradores del Instituto de Tecnologías del Lenguaje de Carnegie Mellon University (CMU) en el uso del software libre de captura y transcripción de voz, CMU Sphinx versión 3 (desarrollado por ellos). El software fue adaptado para generar un modelo de lenguaje Castellano chileno, para lo cual se le incorporó información de fonemas asociados a nuestras formas locales de pronunciación y vocabulario, fundamentalmente asociado a las clases de matemáticas.
El identificador automático de contenidos y prácticas fue construido con el algoritmo de aprendizaje automático Random Forest [4]. Para esto, primero a cada texto se le eliminan las palabras que no representan contenido, sino más bien juegan un rol formal en el lenguaje, tales como artículos, pronombres, conectores, etc., y luego se agrupan las palabras según su raíz. Posteriormente, para cada categoría de clasificación, el algoritmo Random Forest cuenta cuántas veces se repiten las diferentes raíces en cada texto por separado, y a partir de las clasificaciones manuales, le asigna a cada raíz de cada texto un peso (importancia), generando así una bolsa de palabras que se asocian a cada categoría de clasificación.

3.2. Resultados

El desempeño de un clasificador automático depende de su capacidad de replicar las identificaciones realizadas por el/los experto(s) en el conjunto de testeo. Para evaluar la clasificación automática, se utilizó el indicador Alternative Chance-Corrected Coefficient (AC1) [5], introducido en el capítulo previo, que mide el nivel de concordancia entre los expertos y el clasificador automático, tomando valores menores a 1, de modo que valores cercanos a 1 indican un mayor nivel de concordancia o acuerdo entre los evaluadores.
Como fue mencionado anteriormente, las primeras 309 clases fueron clasificadas por 3 expertos cada una, y las restantes clases fueron clasificadas por una o más personas. A raíz de esto, para los textos que clasificaron los tres expertos, primero se entrenó el sistema con la clasificación de cada experto por separado, utilizando 214 clases para entrenar y 95 para evaluar el nivel de acuerdo. Los niveles de concordancia entre los expertos y el clasificador automático (AC1) de los 3 entrenamientos se muestran en las figuras 3.1 y 3.2.
En la siguiente etapa se consolidaron las clasificaciones manuales de los tres expertos de manera tal que: (1) si dos o más expertos dicen que una categoría está presente en un texto, entonces se considera presente, mientras que (2) si solamente 1 o ninguno de los 3 dice que está presente, entonces no lo está. Así, se genera una 'Clasificación Experta Consolidada', para usarla en el entrenamiento del sistema y evaluar si esto produce un mejor desempeño de clasificación del sistema al compararlo con los expertos. En las figuras 3.1 y 3.2 se muestra el AC1 obtenido del entrenamiento del sistema con esta consolidación de los juicios expertos, junto a los asociados a cada experto por separado.
El proyecto en su estado actual contempla continuar entrenando el sistema con todos los textos disponibles en la marcha continuada, aunque no estuvieran clasificados por tres expertos cada uno. Para esto se considera una categoría como presente en un texto si más de la mitad de los clasificadores de dicho texto considera que la categoría en cuestión está presente (voto de mayoría). Los niveles de concordancia entre la Base Final de Entrenamiento y el clasificador automático, con 866 textos clasificados por 53 personas en total, se presentan en la figura 3.3 a continuación. Por último, utilizando también el indicador AC1, se calcula el

acuerdo entre los tres Expertos al clasificar, de a pares y entre los tres. En la figura 3.4 se muestran los valores de AC1 obtenidos, junto a los resultados del entrenamiento final.

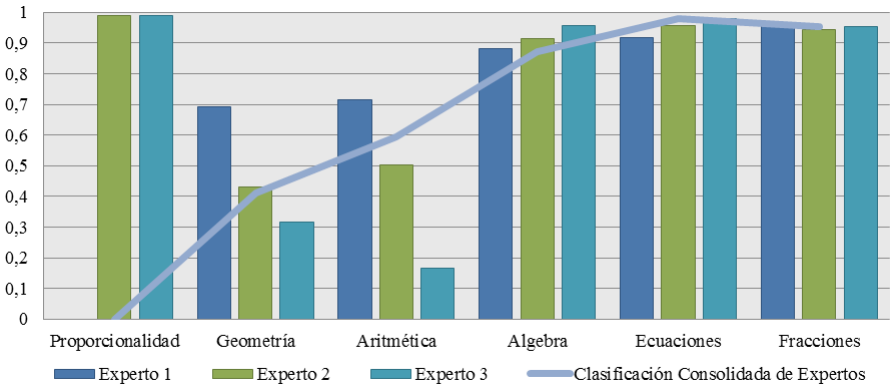

Figura 3.1: AC1 del sistema entrenado con clasificaciones de cada experto y con Clasificación Consolidada - Contenidos

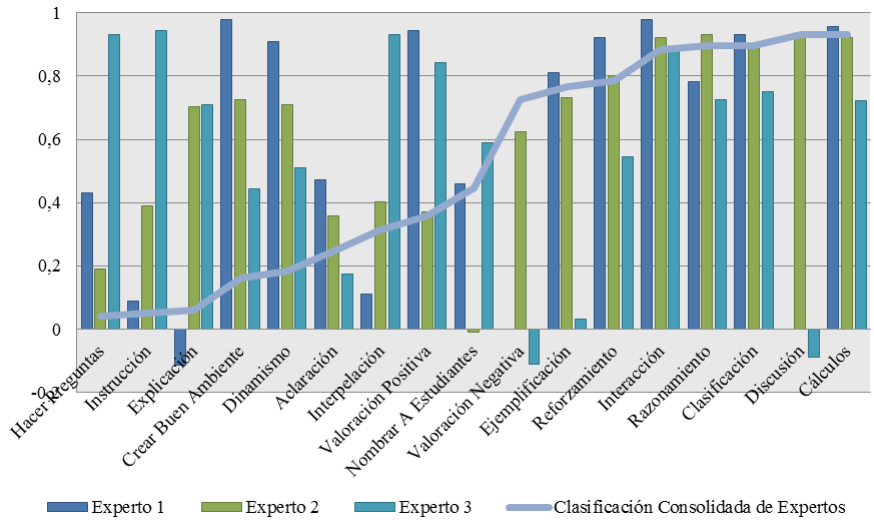

Figura 3.2: AC1 del sistema entrenado con clasificaciones de cada experto y con Clasificación consolidada - Prácticas

3.3. Discusión de los resultados observados

Considerando que no existe, a priori, argumento alguno para preferir el juicio de uno de los expertos en particular frente a los otros, primero se entrena el sistema con cada clasificación por separado para ver el desempeño del clasificador automático según los tres expertos. En las figuras 3.1 y 3.2 se observa que el Experto 1 tiene el mayor AC1 para 12 de las 23 categorías, mientras que los Expertos 2 y 3 tienen el mayor AC1 solamente para 3 y 7 categorías respectivamente (y hay una categoría donde los Expertos 2 y 3 tienen el mismo

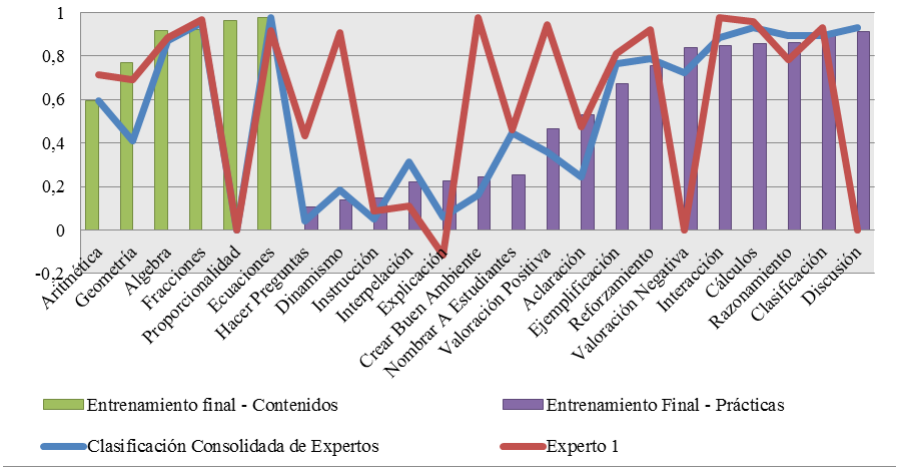

Figura 3.3: AC1 del sistema entrenado con base de entrenamiento final, con clasificación consolidada de 3 expertos y con clasificación de experto 1

Figura 3.4: Acuerdo (AC1) entre expertos y AC1 del sistema entrenado con Base de Entrenamiento Final

AC1). Por otra parte, se calculó el acuerdo entre los expertos. En los Contenidos, el Experto 1 con el Experto 2 están tan de acuerdo como el Experto 2 con el 3, y en las prácticas, se observa una gran variación en los valores encontrados, habiendo incluso varios valores negativos y otros que están bajo 0.4. Lo anterior se debe a la dificultad que conlleva definir exactamente qué es y cómo detectar una práctica en una clase, y por ende, los expertos inevitablemente interpretan las prácticas de acuerdo a su experiencia personal y no de forma completamente objetiva. Para tratar de lograr un desempeño más estable e independiente de quién sea el experto que clasifica las clases, se decidió entrenar el sistema con la 'Clasificación Experta Consolidada'. A partir de la figura 3.1 se observa que el entrenamiento de la clasificación de contenidos con la clasificación consolidada de los tres expertos tiene un valor de AC1 promedio similar al obtenido como media de los AC1 obtenidos con el entrenamiento de los tres Expertos por separado, y en cuanto a las prácticas, la figura 3.2 muestra que el

valor del AC1 del entrenamiento con la Clasificación Consolidada de los tres expertos es más bajo que el obtenido al entrenar el sistema con los Expertos 1, 2 y 3 por separado, pero al utilizar la Clasificación Experta Consolidada, estamos entrenando el sistema con datos más robustos y objetivos, dada la subjetividad en la detección de prácticas docentes en las clases. Por consiguiente, dado que los tres expertos clasificaron solamente una parte de las clases disponibles para entrenar el sistema, se decide seguir entrenando el sistema con todos los textos disponibles, aunque no estuvieran clasificados por tres expertos cada uno. Lo anterior aumenta los datos de entrenamiento de 214 a 624, y los de prueba de 95 a 242. Así, de acuerdo a los resultados presentados en la figura 3.3, el clasificador automático mejoró significativamente su desempeño, sobre todo en aquellas categorías más difíciles de detectar. En particular, considerando que, hasta el momento el mejor desempeño se había obtenido al entrenar el sistema con las clasificaciones del Experto 1, de la figura 3.3 se observa que el entrenamiento final tiene mejor AC1 que el Experto 1 en 11 categorías; y mejor AC1 en 12 categorías respecto a la Clasificación Consolidada de los tres Expertos. Lo anterior es importante porque el sistema en su etapa de producción se deberá seguir reentrenando constantemente a partir de la nueva información que los docentes entregan al utilizar la aplicación grabando y clasificando más clases, y es inviable que un Experto se mantenga de manera permanente encargado de clasificar todas las nuevas clases que se espera ingresar a futuro. Por ende, saber que el clasificador automático obtiene mejores resultados al entrenarlo con clasificaciones realizadas por múltiples personas utilizando la aplicación, hace que este proyecto sea sostenible en el tiempo. Por otra parte, a partir de la figura 3.4 se observa que el sistema entrenado puede clasificar automáticamente los textos y lograr una concordancia con las clasificaciones de los expertos, igual o mejor a la concordancia que logran los expertos entre sí. Dado que la base final de entrenamiento es creada a partir del promedio de los acuerdos entre expertos clasificadores, no es realista pensar que un autómata pueda estar más de acuerdo con los expertos que los expertos entre ellos, y por ende un buen resultado es aquel que se asemeja al acuerdo que logran los expertos entre sí, porque indica que el sistema se puede comportar como un experto más en la clasificación. Al observar los resultados del entrenamiento final de la figura 3.3, de las 23 categorías a clasificar, 5 tienen un valor de AC1 sobre 0.9, 10 están sobre 0.8 y 16 de las 23 categorías están sobre 0.4. Al separar por categorías de Contenidos y Prácticas, tenemos que 4 de los 6 Contenidos están sobre 0.9 y las 6 categorías de contenidos están sobre 0.55. Por último, 8 de las 17 prácticas superan 0.6 y 10 superan 0.4. Estos resultados muestran que sí es posible clasificar tanto contenidos como prácticas de forma automática a partir de un texto. Los resultados son mejores en la identificación de Contenidos que Prácticas. Esto se explica por el hecho de que, a diferencia de un contenido, la presencia o no de una práctica docente puede manifestarse desde muchas dimensiones (por ejemplo, tono de voz, gestos, etc.), siendo el texto hablado solo una de ellas, es probable que se necesite mucho más entrenamiento del sistema para reconocer prácticas docentes con los mismos niveles de AC1 (acuerdo con expertos) que los contenidos, usando solo texto. Asimismo, dada la mejoría en el desempeño al casi triplicar la cantidad de datos de entrenamiento, esperamos que a mayor cantidad de registros de audios, mejor será el desempeño del clasificador automático.

Conclusión

En este trabajo se ha presentado una síntesis de los resultados conocidos hasta el momento en la utilización de un algoritmo de clasificación de textos correspondientes a transcripciones de clases de matemática, además de presentar una discusión acerca de los métodos utilizados para la evaluación del desempeño del clasificador automatizado al compararlo con el estándar aceptado en la actualidad para la misma tarea, el cual corresponde al juicio de expertos que son capaces de usar su criterio adquirido a través de su experiencia para tomar una decisión. Los resultados son promisorios, dado que el clasificador se comporta como un experto más en el grado de acuerdo alcanzado por el juicio de varios expertos en el tema, es decir, alcanza niveles de acuerdo o confiabilidad de nivel similar al que alcanzan expertos entre sí. Sin embargo, en varias categorías este resultado es aún imperfecto, pues la detección de una determinada práctica docente en algunos casos es controversial entre los expertos humanos, algo que corresponde a una propidad indeseable si lo que se desea es un diagnóstico certero e independiente del observador.

Esta es una observación interesante de hacer, pero probablemente su solución debe provenir de una mejor definición de criterios de evaluación o clasificación para los expertos humanos, de manera de poseer datos de entrenamiento más consistentes antes de intentar entrenar un sistema automatizado, dado que la falta de acuerdo entre los expertos se convierte en ruido que debilita la posibilidad de exhibir amplio acuerdo entre el sistema y un evaluador humano, pues el evaluador automatizado puede tener un criterio uniforme a partir de los datos de entrenamiento únicamente si estos datos son consistentes a través de los expertos. Sin embargo, se puede considerar que el comportamiento actual del clasificador puede llegar a ser una virtud en la aplicación que se busca hacer del sistema, donde los datos para entrenamiento y retroalimentación para el fortalecimiento del sistema provienen de los usuarios de éste, por lo que no poseen la uniformidad tan deseable en un contexto experimental para fijar parámetros de comparación.

El desempeño del coeficiente $AC1$ como indicador del acuerdo que el sistema alcanza con los expertos humanos ha sido satisfactorio y da lugar a información útil, no obstante es importante hacer una crítica y destacar que un punto a mejorar a futuro es recordar que este valor está sujeto a errores muestrales y podría ser de vital importancia en la continuación de este trabajo o en trabajos futuros realizados en un marco similar el hacer un estudio inferencial de los valores de este coeficiente.

Bibliografía

[1] Araya, R.; F. Plana; Dartnell, P.; Soto-Andrade, J.; G. Luci; E. Salinas; M. Araya.2011. Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm *British Journal of Educational Technology*, 43 (6), 837 - 846.

[2] Arnold, Taylor B.; Emerson, John W.. 2011. Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions *The R Journal*, 3 (2), 34 - 39.

[3] Blood E, Spratt KF. 2007. Disagreement on agreement: two alternative agreement coefficients. *SAS Global Forum 2007.*

[4] Breiman, Leo. 2001. Random Forests. *Machine Learning* 45 (1), 5-32.

[5] Gwet, K. L. 2014. *Handbook of inter-rater reliability* 4° Ed. Gaithersburg. Advanced Analytics LLC.

[6] Gwet, K. L. 2008. Computing inter rater reliability and its variance in the pressence of high agreement. *The British Journal of Mathematical and Statistichal Psychology*, 61 (1), 29-48.

[7] Koedinger, K; Booth, J.; Klahr, D. 2013. Instructional Complexity and the Science to Constrain it. *Science Magazine*, 342, 935-937

[8] National Board Resource Center. 2010. A quality teacher in every classroom. Stanford, California: Stanford University.

[9] National Education Association. 1946. Research Bulletin. December 1946, pp. 146-148.

[10] Stevens, R. 1912. The question as a measure of efficiency in instruction. *New York: Bureau of publications*, Teachers College, Columbia University. pp 11, 15-17.